

Financial Econometrics Notes

**Kevin Sheppard
University of Oxford**

November 9, 2020

This version: 11:12, November 9, 2020

©2020 Kevin Sheppard

Contents

1 Probability, Random Variables and Expectations	1
1.1 Axiomatic Probability	1
1.2 Univariate Random Variables	7
1.3 Multivariate Random Variables	20
1.4 Expectations and Moments	33
2 Estimation, Inference, and Hypothesis Testing	57
2.1 Estimation	57
2.2 Convergence and Limits for Random Variables	67
2.3 Properties of Estimators	71
2.4 Distribution Theory	77
2.5 Hypothesis Testing	94
2.6 The Bootstrap and Monte Carlo	106
2.7 Inference on Financial Data	110
3 Analysis of Cross-Sectional Data	127
3.1 Model Description	127
3.2 Functional Form	130
3.3 Estimation	133
3.4 Assessing Fit	136
3.5 Assumptions	141
3.6 Small-Sample Properties of OLS estimators	142
3.7 Maximum Likelihood	144
3.8 Small-Sample Hypothesis Testing	146
3.9 Large-Sample Assumption	161
3.10 Large-Sample Properties	162
3.11 Large-Sample Hypothesis Testing	164
3.12 Violations of the Large-Sample Assumptions	169
3.13 Model Selection and Specification Checking	182
3.14 Machine Learning	197
3.15 Projection	208
3.A Selected Proofs	210

4 Analysis of a Single Time Series	225
4.1 Stochastic Processes	225
4.2 Stationarity, Ergodicity, and the Information Set	225
4.3 ARMA Models	228
4.4 Difference Equations	236
4.5 Data and Initial Estimates	244
4.6 Autocorrelations and Partial Autocorrelations	245
4.7 Estimation	256
4.8 Inference	259
4.9 Forecasting	260
4.10 Nonstationary Time Series	265
4.11 Nonlinear Models for Time-Series Analysis	276
4.12 Filters	276
4.A Computing Autocovariance and Autocorrelations	290
5 Analysis of Multiple Time Series	317
5.1 Vector Autoregressions	317
5.2 Companion Form	323
5.3 Empirical Examples	323
5.4 VAR forecasting	326
5.5 Estimation and Identification	328
5.6 Granger causality	333
5.7 Impulse Response Functions	334
5.8 Cointegration	339
5.9 Cross-sectional Regression with Time-series Data	358
5.A Cointegration in a trivariate VAR	364
6 Generalized Method Of Moments (GMM)	377
6.1 Classical Method of Moments	377
6.2 Examples	378
6.3 General Specification	382
6.4 Estimation	385
6.5 Asymptotic Properties	390
6.6 Covariance Estimation	394
6.7 Special Cases of GMM	398
6.8 Diagnostics	403
6.9 Parameter Inference	404
6.10 Two-Stage Estimation	407
6.11 Weak Identification	409
6.12 Considerations for using GMM	410
7 Univariate Volatility Modeling	413

7.1	Why does volatility change?	413
7.2	ARCH Models	415
7.3	Estimation and Inference	432
7.4	GARCH-in-Mean	438
7.5	Alternative Distributional Assumptions	438
7.6	Model Building	441
7.7	Forecasting Volatility	444
7.8	Realized Variance	450
7.9	Implied Volatility and VIX	458
7.A	Kurtosis of an ARCH(1)	465
7.B	Kurtosis of a GARCH(1,1)	468
8	Value-at-Risk, Expected Shortfall and Density Forecasting	479
8.1	Defining Risk	479
8.2	Value-at-Risk (VaR)	480
8.3	Conditional Value-at-Risk	482
8.4	Unconditional Value at Risk	490
8.5	Evaluating VaR models	492
8.6	Expected Shortfall	497
8.7	Density Forecasting	498
8.8	Coherent Risk Measures	508
9	Multivariate Volatility, Dependence and Copulas	517
9.1	Introduction	517
9.2	Preliminaries	518
9.3	Simple Models of Multivariate Volatility	521
9.4	Multivariate ARCH Models	529
9.5	Realized Covariance	540
9.6	Measuring Dependence	550
9.7	Copulas	558
9.A	Bootstrap Standard Errors	570

List of Figures

1.1	Set Operations	3
1.2	Bernoulli Random Variables	9
1.3	Normal pdf and cdf	12
1.4	Poisson and χ^2 distributions	14
1.5	Bernoulli Random Variables	16
1.6	Joint and Conditional Distributions	27
1.7	Joint distribution of the FTSE 100 and S&P 500	31
1.8	Simulation and Numerical Integration	34
1.9	Modes	39
2.1	Convergence in Distribution	68
2.2	Consistency and Central Limits	75
2.3	Central Limit Approximations	76
2.4	Data Generating Process and Asymptotic Covariance of Estimators	89
2.5	Power	98
2.6	Standard Normal cdf and Empirical cdf	107
2.7	CRSP Value Weighted Market (VWM) Excess Returns	113
3.1	Rejection regions of a t_{10}	149
3.2	Bivariate F distributions	151
3.3	Rejection region of a $F_{5,30}$ distribution	153
3.4	Location of the three test statistic statistics	161
3.5	Effect of correlation on the variance of $\hat{\beta}^{IV}$	176
3.6	Gains of using GLS	181
3.7	Neglected Nonlinearity and Residual Plots	186
3.8	Rolling Parameter Estimates in the 4-Factor Model	189
3.9	Recursive Parameter Estimates in the 4-Factor Model	190
3.10	Influential Observations	192
3.11	Correct and Incorrect use of “Robust” Estimators	196
3.12	The left panel shows the ridge regression restriction for a specific value of ω along with three lines that trace combinations of β_1 and β_2 that produce the same model SSE. The ridge estimate is defined as the point where the SSE is just tangent to a restriction. The right shows the LASSO constraint along with the iso-SSE curves for the same data generating process.	203

3.13 The top panel shows the path of the ridge regression estimates from the four factor model $BH^e = \beta_1 + \beta_2 VWM_i^e + \beta_3 SMB_i + \beta_4 HML_i + \beta_5 MOM_i + \varepsilon_i$. The penalty parameter ω is increased from zero to the value that produces the OLS estimate. The bottom panel contains the path of the LASSO estimates as the restriction is decreased. The kinks indicate points where a parameter switches from being exactly zero to a non-zero value.	204
3.14 A regression tree where the left-hand-side variable is the return on the Big-High portfolio and the model is built using the four factors: VWM^e , SMB , HML , and MOM . The first and second splits used the market portfolio to bin the returns into four regions ranging from very low to very high. The final level splits used different variables so that the terminal leaves depend on both the market and the size factor.	205
3.15 The regression tree implied by the first two splits and the OLS fit of the excess returns on the Big-High portfolio on the market.	206
3.16 Weights of an S&P 500 Tracking Portfolio	209
4.1 Dynamics of linear difference equations	242
4.2 Stationarity of an AR(2)	244
4.3 VWM and Default Spread	245
4.4 ACF and PACF for ARMA Processes	250
4.5 ACF and PACF for ARMA Processes	251
4.6 Autocorrelations and Partial Autocorrelations for the VWM and the Default Spread	255
4.7 M1, M1 growth, and the ACF and PACF of M1 growth	267
4.8 Time Trend Models of GDP	269
4.9 Unit Root Analysis of $\ln CPI$ and the Default Spread	275
4.10 Ideal Filters	278
4.11 Actual Filters	282
4.12 Cyclical Component of U.S. Real GDP	285
4.13 Markov Switching Processes	289
4.14 Self Exciting Threshold Autoregression Processes	291
4.15 Plots for question 2(b).	307
4.16 Exercise 4.9	311
4.17 Plots for question 2(b).	315
5.1 Simulated data from a VAR(22)	329
5.2 Vector ACF and CCF	330
5.3 Vector PACF and PCCF	331
5.4 Impulse Response Functions	337
5.5 Cointegration	341
5.6 Detrended cay Residuals	357
5.7 Impulse Response of Level-Slope-Curvature	371
6.1 2-Step GMM Objective Function Surface	390
7.1 Returns of the S&P 500 and WTI	420
7.2 Squared returns of the S&P 500 and WTI	421

7.3	Absolute returns of the S&P 500 and WTI	428
7.4	News impact curves	431
7.5	Various estimated densities for the S&P 500	440
7.6	Effect of distribution on volatility estimates	442
7.7	ACF and PACF of S&P 500 squared returns	444
7.8	ACF and PACF of WTI squared returns	445
7.9	RV^{AC1} and sampling frequency	454
7.10	Volatility Signature Plot for SPY RV	455
7.11	Realized Variance and sampling frequency	458
7.12	Black-Scholes Implied Volatility	460
7.13	Option prices generated from the Black-Scholes pricing formula for an underlying with a price of \$100 with a volatility of 20% or 60% (bottom). The options expire in 1 month ($T = 1/12$), and the risk-free rate is 2%. The solid lines show the out-of-the-money options that are used to compute the VIX. The solid markers show the values where the option price to be at least \$0.01 using a \$4 grid of strike prices.	464
7.14	VIX and alternative measures of volatility	466
7.15	Plots for question 7.16.	478
8.1	Graphical representation of Value-at-Risk	481
8.2	Estimated % VaR for the S&P 500	488
8.3	The estimated 5% VaR for the S&P 500 using weighted Historical Simulation for $\lambda \in \{0.95, 0.99, 0.999\}$. The three values of λ place 90% of the weight on the most recent 45, 230, and 2280 observations, respectively. Larger values of the decay parameter λ produce smoother conditional VaR estimates.	489
8.4	S&P 500 Returns and a Parametric Density	493
8.5	Empirical and Smoothed empirical Distribution and Density	501
8.6	Naïve and Correct Density Forecasts	502
8.7	Fan plot	503
8.8	QQ plot	504
8.9	Kolmogorov-Smirnov plot	506
8.10	Returns, Historical Simulation VaR and Normal GARCH VaR.	513
9.1	Lag weights in RiskMetrics methodologies	523
9.2	Rolling Window Correlation Measures	530
9.3	Observable and Principal Component Correlation Measures	531
9.4	Volatility from Multivariate Models	541
9.5	Small-Cap - Large-Cap Correlation	542
9.6	Small-Cap - Long Government Bond Correlation	543
9.7	Large-Cap - Bond Correlation	544
9.8	ETF Transactions per Day	548
9.9	Pseudo-correlation and Cross-volatility Signatures	551
9.10	ETF Realized Correlation	552
9.11	Rolling Dependence Measures	555
9.12	Exceedance Correlation	556
9.13	Symmetric and Asymmetric Dependence	559
9.14	Copula Distributions and Densities	566

9.15 Copula Densities with Standard Normal Margins	567
9.16 S&P 500 - FTSE 100 Diagnostics	571
9.17 S&P 500 and FTSE 100 Exceedance Correlations	576

List of Tables

1.1	Monte Carlo and Numerical Integration	55
2.1	Parameter Values of Mixed Normals	90
2.2	Outcome matrix for a hypothesis test	96
2.3	Inference on the Market Premium	112
2.4	Inference on the Market Premium	112
2.5	Comparing the Variance of the NASDAQ and S&P 100	114
2.6	Comparing the Variance of the NASDAQ and S&P 100	116
2.7	Wald, LR and LM Tests	122
3.1	Fama-French Data Description	130
3.2	Descriptive Statistics of the Fama-French Data Set	131
3.3	Regression Coefficient on the Fama-French Data Set	136
3.4	Centered and Uncentered R^2 and \bar{R}^2 with Regressor Changes	140
3.5	t -stats for the Big-High Portfolio	155
3.6	Likelihood Ratio Tests on the Big-High Portfolio	157
3.7	Comparison of Small- and Large- Sample t -Statistics	169
3.8	Comparison of Small- and Large- Sample Wald, LR, and LM Statistic	170
3.9	OLS and GLS Parameter Estimates and t -stats	182
4.1	Estimates from Time-Series Models	245
4.2	ACF and PACF for ARMA processes	249
4.3	Seasonal Model Estimates	268
4.4	Unit Root Analysis of $\ln CPI$	276
5.1	Parameter estimates from A Monetary Policy VAR	325
5.2	Relative out-of-sample Mean Square Error for forecasts between 1 and 8-quarters ahead. The benchmark model is a constant for the unemployment rate and the inflation rate and a random walk for the Federal Funds rate. Model parameters are recursively estimated, and forecasts are produced once 50% of the available sample. Model order is selected using the BIC.	327
5.3	AIC and BIC in a Monetary Policy VAR	332
5.4	Granger Causality	334
5.5	Johansen Methodology	352
5.6	Unit Root and Cointegration Tests	355
5.7	Comparing Engle-Granger and Dynamic OLS	355

6.1	Parameter Estimates from a Consumption-Based Asset Pricing Model	389
6.2	Stochastic Volatility Model Parameter Estimates	391
6.3	Effect of Covariance Estimator on GMM Estimates	397
6.4	Stochastic Volatility Model Monte Carlo	399
6.5	Tests of a Linear Factor Model	404
6.6	Fama-MacBeth Inference	409
7.1	Summary statistics for the S&P 500 and WTI	426
7.2	Parameter estimates from ARCH-family models	427
7.3	Bollerslev-Wooldridge Covariance estimates	436
7.4	GARCH-in-mean estimates	439
7.5	Model selection for the S&P 500	443
7.6	Model selection for WTI	446
7.7	Option prices generated from the Black-Scholes pricing formula for an underlying with a price of \$100 with a volatility of 20%. The options expire in 1 month ($T = 1/12$), and the risk-free rate is 2%. The third column shows the absolute difference which is used to determine K_0 in the VIX formula. The final column contains the contribution of each option to the VIX as measured by $\frac{2}{T} \exp(rT) \Delta K_i / K_i^2 \times Q(K_i)$	465
8.1	Estimated model parameters and quantiles	489
8.2	Unconditional VaR of the S&P 500	492
9.1	Principal Component Analysis of the S&P 500	526
9.2	Correlation Measures for the S&P 500	529
9.3	CCC GARCH Correlation	539
9.4	Multivariate GARCH Model Estimates	540
9.5	Refresh-time sampling	546
9.6	Dependence Measures for Weekly FTSE and S&P 500 Returns	554
9.7	Copula Tail Dependence	565
9.8	Unconditional Copula Estimates	569
9.9	Conditional Copula Estimates	570

Chapter 1

Probability, Random Variables and Expectations

Note: The primary reference for these notes is Mittelhammer (1999). Other treatments of probability theory include Gallant (1997), Casella and Berger (2001) and Grimmett and Stirzaker (2001).

This chapter provides an overview of probability theory as it applied to both discrete and continuous random variables. The material covered in this chapter serves as a foundation of the econometric sequence and is useful throughout financial economics. The chapter begins with a discussion of the axiomatic foundations of probability theory and then proceeds to describe properties of univariate random variables. Attention then turns to multivariate random variables and important difference from univariate random variables. Finally, the chapter discusses the expectations operator and moments.

1.1 Axiomatic Probability

Probability theory is derived from a small set of axioms – a minimal set of essential assumptions. A deep understanding of axiomatic probability theory is *not* essential to financial econometrics or to the use of probability and statistics in general, although understanding these core concepts does provide additional insight.

The first concept in probability theory is the sample space, which is an abstract concept containing primitive probability events.

Definition 1.1 (Sample Space). The sample space is a set, Ω , that contains all possible outcomes.

Example 1.1. Suppose interest is on a standard 6-sided die. The sample space is 1-dot, 2-dots, ..., 6-dots.

Example 1.2. Suppose interest is in a standard 52-card deck. The sample space is then $A\clubsuit, 2\clubsuit, 3\clubsuit, \dots, J\clubsuit, Q\clubsuit, K\clubsuit, A\diamondsuit, \dots, K\diamondsuit, A\heartsuit, \dots, K\heartsuit$.

Example 1.3. Suppose interest is in the logarithmic stock return, defined as $r_t = \ln P_t - \ln P_{t-1}$, then the sample space is \mathbb{R} , the real line.

The next item of interest is an event.

Definition 1.2 (Event). An event, ω , is a subset of the sample space Ω .

An event may be any subsets of the sample space Ω (including the entire sample space), and the set of all events is known as the event space.

Definition 1.3 (Event Space). The set of all events in the sample space Ω is called the event space, and is denoted \mathcal{F} .

Event spaces are a somewhat more difficult concept. For finite event spaces, the event space is usually the power set of the outcomes – that is, the set of all possible unique sets that can be constructed from the elements. When variables can take infinitely many outcomes, then a more nuanced definition is needed, although the main idea is to define the event space to be all non-empty intervals (so that each interval has infinitely many points in it).

Example 1.4. Suppose interest lies in the outcome of a coin flip. Then the sample space is $\{H, T\}$ and the event space is $\{\emptyset, \{H\}, \{T\}, \{H, T\}\}$ where \emptyset is the empty set.

The first two axioms of probability are simple: all probabilities must be non-negative and the total probability of all events is one.

Axiom 1.1. For any event $\omega \in \mathcal{F}$,

$$\Pr(\omega) \geq 0. \quad (1.1)$$

Axiom 1.2. The probability of all events in the sample space Ω is unity, i.e.

$$\Pr(\Omega) = 1. \quad (1.2)$$

The second axiom is a normalization that states that the probability of the entire sample space is 1 and ensures that the sample space must contain all events that may occur. $\Pr(\cdot)$ is a set-valued function – that is, $\Pr(\omega)$ returns the probability, a number between 0 and 1, of observing an event ω .

Before proceeding, it is useful to refresh four concepts from set theory.

Definition 1.4 (Set Union). Let A and B be two sets, then the union is defined

$$A \cup B = \{x : x \in A \text{ or } x \in B\}.$$

A union of two sets contains all elements that are in either set.

Definition 1.5 (Set Intersection). Let A and B be two sets, then the intersection is defined

$$A \cap B = \{x : x \in A \text{ and } x \in B\}.$$

The intersection contains only the elements that are in both sets.

Definition 1.6 (Set Complement). Let A be a set, then the complement set, denoted

$$A^c = \{x : x \notin A\}.$$

The complement of a set contains all elements which are not contained in the set.

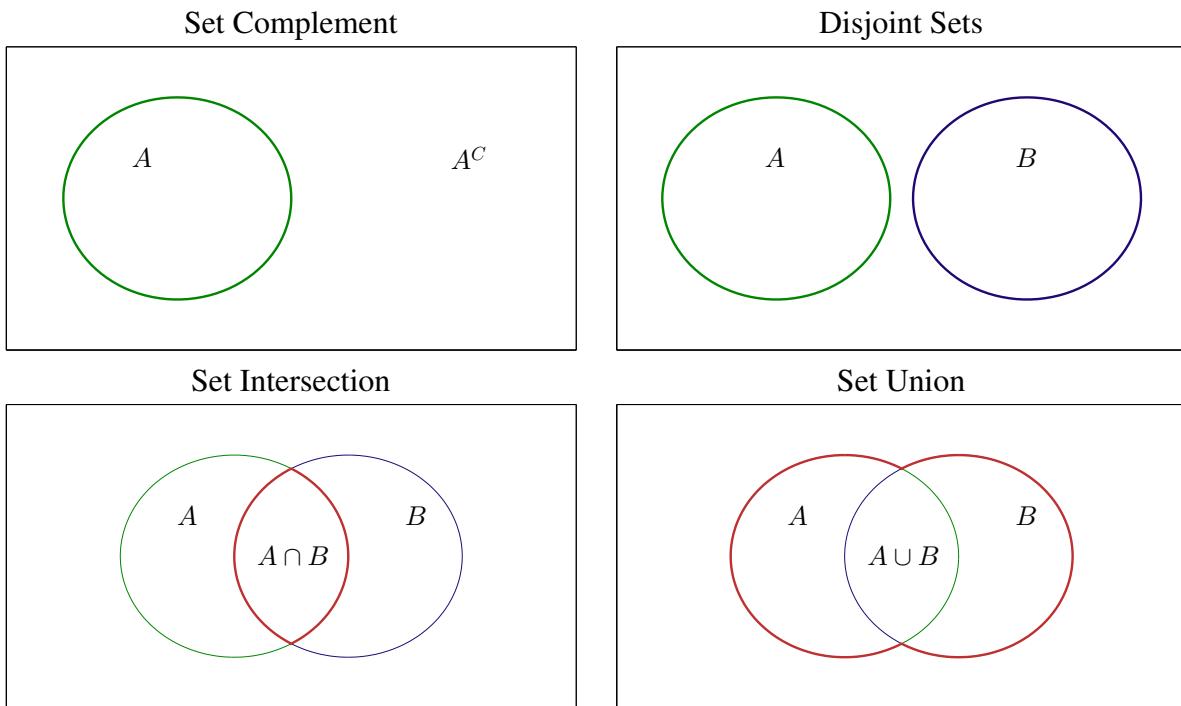


Figure 1.1: The four set definitions presented in \mathbb{R}^2 . The upper left panel shows a set and its complement. The upper right shows two disjoint sets. The lower left shows the intersection of two sets (darkened region) and the lower right shows the union of two sets (darkened region). In all diagrams, the outer box represents the entire space.

Definition 1.7 (Disjoint Sets). Let A and B be sets, then A and B are disjoint if and only if $A \cap B = \emptyset$.

Figure 1.1 provides a graphical representation of the four set operations in a 2-dimensional space. The third and final axiom states that probability is additive when sets are disjoint.

Axiom 1.3. Let $\{A_i\}$, $i = 1, 2, \dots$ be a finite or countably infinite set of disjoint events.¹ Then

$$\Pr \left(\bigcup_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} \Pr(A_i). \quad (1.3)$$

Assembling a sample space, event space and a probability measure into a set produces what is known as a probability space. Throughout the course, and in virtually all statistics, a complete probability space is assumed (typically without explicitly stating this assumption).²

¹

Definition 1.8. A S set is countably infinite if there exists a bijective (one-to-one) function from the elements of S to the natural numbers $\mathbb{N} = \{1, 2, \dots\}$. Common sets that are countable infinite include the integers (\mathbb{Z}) and the rational numbers (\mathbb{Q}).

²A complete probability space is complete if and only if $B \in \mathcal{F}$ where $\Pr(B) = 0$ and $A \subset B$, then $A \in \mathcal{F}$. This condition ensures that probability can be assigned to any event.

Definition 1.9 (Probability Space). A probability space is denoted using the tuple $(\Omega, \mathcal{F}, \Pr)$ where Ω is the sample space, \mathcal{F} is the event space and \Pr is the probability set function which has domain $\omega \in \mathcal{F}$.

The three axioms of modern probability are very powerful, and a large number of theorems can be proven using only these axioms. A few simple examples are provided, and selected proofs appear in the Appendix.

Theorem 1.1. *Let A be an event in the sample space Ω , and let A^c be the complement of A so that $\Omega = A \cup A^c$. Then $\Pr(A) = 1 - \Pr(A^c)$.*

Since A and A^c are disjoint, and by definition A^c is everything not in A , then the probability of the two must be unity.

Theorem 1.2. *Let A and B be events in the sample space Ω . Then $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$.*

This theorem shows that for any two sets, the probability of the union of the two sets is equal to the probability of the two sets minus the probability of the intersection of the sets.

1.1.1 Conditional Probability

Conditional probability extends the basic concepts of probability to the case where interest lies in the probability of one event conditional on the occurrence of another event.

Definition 1.10 (Conditional Probability). Let A and B be two events in the sample space Ω . If $\Pr(B) \neq 0$, then the conditional probability of the event A , given event B , is given by

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}. \quad (1.4)$$

The definition of conditional probability is intuitive. The probability of observing an event in set A , given an event in the set B has occurred, is the probability of observing an event in the intersection of the two sets normalized by the probability of observing an event in set B .

Example 1.5. In the example of rolling a die, suppose $A = \{1, 3, 5\}$ is the event that the outcome is odd and $B = \{1, 2, 3\}$ is the event that the outcome of the roll is less than 4. Then the conditional probability of A given B is

$$\frac{\Pr(\{1, 3\})}{\Pr(\{1, 2, 3\})} = \frac{\frac{2}{6}}{\frac{3}{6}} = \frac{2}{3}$$

since the intersection of A and B is $\{1, 3\}$.

The axioms can be restated in terms of conditional probability, where the sample space consists of the events in the set B .

1.1.2 Independence

Independence of two measurable sets means that any information about an event occurring in one set has no information about whether an event occurs in another set.

Definition 1.11. Let A and B be two events in the sample space Ω . Then A and B are independent if and only if

$$\Pr(A \cap B) = \Pr(A)\Pr(B) \quad (1.5)$$

, $A \perp\!\!\!\perp B$ is commonly used to indicate that A and B are independent.

One immediate implication of the definition of independence is that when A and B are independent, then the conditional probability of one given the other is the same as the unconditional probability of the random variable – i.e. $\Pr(A|B) = \Pr(A)$.

1.1.3 Bayes Rule

Bayes rule is frequently encountered in both statistics (known as Bayesian statistics) and in financial models where agents learn about their environment. Bayes rule follows as a corollary to a theorem that states that the total probability of a set A is equal to the conditional probability of A given a set of disjoint sets B which span the sample space.

Theorem 1.3. Let $B_i, i = 1, 2, \dots$ be a finite or countably infinite partition of the sample space Ω so that $B_j \cap B_k = \emptyset$ for $j \neq k$ and $\bigcup_{i=1}^{\infty} B_i = \Omega$. Let $\Pr(B_i) > 0$ for all i , then for any set A ,

$$\Pr(A) = \sum_{i=1}^{\infty} \Pr(A|B_i) \Pr(B_i). \quad (1.6)$$

Bayes rule restates the previous theorem so that the probability of observing an event in B_j given an event in A is observed can be related to the conditional probability of A given B_j .

Corollary 1.1 (Bayes Rule). Let $B_i, i = 1, 2, \dots$ be a finite or countably infinite partition of the sample space Ω so that $B_j \cap B_k = \emptyset$ for $j \neq k$ and $\bigcup_{i=1}^{\infty} B_i = \Omega$. Let $\Pr(B_i) > 0$ for all i , then for any set A where $\Pr(A) > 0$,

$$\begin{aligned} \Pr(B_j|A) &= \frac{\Pr(A|B_j) \Pr(B_j)}{\sum_{i=1}^{\infty} \Pr(A|B_i) \Pr(B_i)}. \\ &= \frac{\Pr(A|B_j) \Pr(B_j)}{\Pr(A)} \end{aligned}$$

An immediate consequence of the definition of conditional probability is the

$$\Pr(A \cap B) = \Pr(A|B) \Pr(B),$$

which is referred to as the multiplication rule. Also notice that the order of the two sets is arbitrary, so that the rule can be equivalently stated as $\Pr(A \cap B) = \Pr(B|A) \Pr(A)$. Combining these two (as long as $\Pr(A) > 0$),

$$\begin{aligned} \Pr(A|B) \Pr(B) &= \Pr(B|A) \Pr(A) \\ \Rightarrow \Pr(B|A) &= \frac{\Pr(A|B) \Pr(B)}{\Pr(A)}. \end{aligned} \quad (1.7)$$

Example 1.6. Suppose a family has 2 children and one is a boy, and that the probability of having a child of either sex is equal and independent across children. What is the probability that they have 2 boys?

Before learning that one child is a boy, there are 4 equally probable possibilities: $\{B, B\}$, $\{B, G\}$, $\{G, B\}$ and $\{G, G\}$. Using Bayes rule,

$$\begin{aligned}\Pr(\{B, B\} | B \geq 1) &= \frac{\Pr(B \geq 1 | \{B, B\}) \times \Pr(\{B, B\})}{\sum_{S \in \{\{B, B\}, \{B, G\}, \{G, B\}, \{G, G\}\}} \Pr(B \geq 1 | S) \Pr(S)} \\ &= \frac{1 \times \frac{1}{4}}{1 \times \frac{1}{4} + 1 \times \frac{1}{4} + 1 \times \frac{1}{4} + 0 \times \frac{1}{4}} \\ &= \frac{1}{3}\end{aligned}$$

so that knowing one child is a boy increases the probability of 2 boys from $\frac{1}{4}$ to $\frac{1}{3}$. Note that

$$\sum_{S \in \{\{B, B\}, \{B, G\}, \{G, B\}, \{G, G\}\}} \Pr(B \geq 1 | S) \Pr(S) = \Pr(B \geq 1).$$

Example 1.7. The famous Monte Hall *Let's Make a Deal* television program is an example of Bayes rule. Contestants competed for one of three prizes, a large one (e.g. a car) and two uninteresting ones (duds). The prizes were hidden behind doors numbered 1, 2 and 3. Before the contest starts, the contestant has no information about which door has the large prize, and to the initial probabilities are all $\frac{1}{3}$. During the negotiations with the host, it is revealed that one of the non-selected doors does *not* contain the large prize. The host then gives the contestant the chance to switch from the door initially chosen to the one remaining door. For example, suppose the contestant choose door 1 initially, and that the host revealed that the large prize is not behind door 3. The contestant then has the chance to choose door 2 or to stay with door 1. In this example, B is the event where the contestant chooses the door which hides the large prize, and A is the event that the large prize is not behind door 2.

Initially there are three equally likely outcomes (from the contestant's point of view), where D indicates dud, L indicates the large prize, and the order corresponds to the door number.

$$\{D, D, L\}, \{D, L, D\}, \{L, D, D\}$$

The contestant has a $\frac{1}{3}$ chance of having the large prize behind door 1. The host will never remove the large prize, and so applying Bayes rule we have

$$\begin{aligned}\Pr(L = 2 | H = 3, S = 1) &= \frac{\Pr(H = 3 | S = 1, L = 2) \times \Pr(L = 2 | S = 1)}{\sum_{i=1}^3 \Pr(H = 3 | S = 1, L = i) \times \Pr(L = i | S = 1)} \\ &= \frac{1 \times \frac{1}{3}}{\frac{1}{2} \times \frac{1}{3} + 1 \times \frac{1}{3} + 0 \times \frac{1}{3}} \\ &= \frac{\frac{1}{3}}{\frac{1}{2}} \\ &= \frac{2}{3}.\end{aligned}$$

where H is the door the host reveals, S is initial door selected, and L is the door containing the large prize. This shows that the probability the large prize is behind door 2, given that the player initially selected door 1 and the host revealed door 3 can be computed using Bayes rule.

$\Pr(H = 3|S = 1, L = 2)$ is the probability that the host shows door 3 given the contestant selected door 1 and the large prize is behind door 2, which always happens since the host will never reveal the large prize. $P(L = 2|S = 1)$ is the probability that the large is in door 2 given the contestant selected door 1, which is $\frac{1}{3}$. $\Pr(H = 3|S = 1, L = 1)$ is the probability that the host reveals door 3 given that door 1 was selected and contained the large prize, which is $\frac{1}{2}$, and $P(H = 3|S = 1, L = 3)$ is the probability that the host reveals door 3 given door 3 contains the prize, which never happens.

Bayes rule shows that it is always optimal to switch doors. This is a counter-intuitive result and occurs since the host's action reveals information about the location of the large prize. Essentially, the two doors not selected by the host have combined probability $\frac{2}{3}$ of containing the large prize before the doors are opened – opening the third assigns its probability to the door not opened.

1.2 Univariate Random Variables

Studying the behavior of random variables, and more importantly functions of random variables (i.e. statistics) is essential for both the theory and practice of financial econometrics. This section covers univariate random variables and multivariate random variables are discussed later.

The previous discussion of probability is set based and so includes objects which cannot be described as random variables, which are a limited (but highly useful) sub-class of all objects that can be described using probability theory. The primary characteristic of a random variable is that it takes values on the real line.

Definition 1.12 (Random Variable). Let $(\Omega, \mathcal{F}, \Pr)$ be a probability space. If $X : \Omega \rightarrow \mathbb{R}$ is a real-valued function have as its domain elements of Ω , then X is called a random variable.

A random variable is essentially a function which takes $\omega \in \Omega$ as an input and returns a value $x \in \mathbb{R}$, where \mathbb{R} is the symbol for the real line. Random variables come in one of three forms: discrete, continuous and mixed. Random variables which mix discrete and continuous distributions are generally less important in financial economics and so here the focus is on discrete and continuous random variables.

Definition 1.13 (Discrete Random Variable). A random variable is called discrete if its range consists of a countable (possibly infinite) number of elements.

While discrete random variables are less useful than continuous random variables, they are still commonly encountered.

Example 1.8. A random variable which takes on values in $\{0, 1\}$ is known as a Bernoulli random variable, and is the simplest non-degenerate random variable (see Section 1.2.3.1).³ Bernoulli random variables are often used to model “success” or “failure”, where success is loosely defined – a large negative return, the existence of a bull market or a corporate default.

The distinguishing characteristic of a discrete random variable is not that it takes only finitely many values, but that the values it takes are distinct in the sense that it is possible to fit small intervals around each point without the overlap.

³A degenerate random variable always takes the same value, and so is not meaningfully random.

Example 1.9. Poisson random variables take values in $\{0, 1, 2, 3, \dots\}$ (an infinite range), and are commonly used to model hazard rates (i.e. the number of occurrences of an event in an interval). They are especially useful in modeling trading activity (see Section 1.2.3.2).

1.2.1 Mass, Density, and Distribution Functions

Discrete random variables are characterized by a probability mass function (pmf) which gives the probability of observing a particular value of the random variable.

Definition 1.14 (Probability Mass Function). The probability mass function, f , for a discrete random variable X is defined as $f(x) = \Pr(x)$ for all $x \in R(X)$, and $f(x) = 0$ for all $x \notin R(X)$ where $R(X)$ is the range of X (i.e. the values for which X is defined).

Example 1.10. The probability mass function of a Bernoulli random variable takes the form

$$f(x; p) = p^x (1 - p)^{1-x}$$

where $p \in [0, 1]$ is the probability of success.

Figure 1.2 contains a few examples of Bernoulli pmfs using data from the FTSE 100 and S&P 500 over the period 1984–2012. Both weekly returns, using Friday to Friday prices and monthly returns, using end-of-month prices, were constructed. Log returns were used ($r_t = \ln(P_t/P_{t-1})$) in both examples. Two of the pmfs defined success as the return being positive. The other two define the probability of success as a return larger than -1% (weekly) or larger than -4% (monthly). These show that the probability of a positive return is much larger for monthly horizons than for weekly.

Example 1.11. The probability mass function of a Poisson random variable is

$$f(x; \lambda) = \frac{\lambda^x}{x!} \exp(-\lambda)$$

where $\lambda \in [0, \infty)$ determines the intensity of arrival (the average value of the random variable).

The pmf of the Poisson distribution can be evaluated for every value of $x \geq 0$, which is the support of a Poisson random variable. Figure 1.4 shows empirical distribution tabulated using a histogram for the time elapsed where .1% of the daily volume traded in the S&P 500 tracking ETF SPY on May 31, 2012. This data series is a good candidate for modeling using a Poisson distribution.

Continuous random variables, on the other hand, take a continuum of values – technically an uncountable infinity of values.

Definition 1.15 (Continuous Random Variable). A random variable is called continuous if its range is uncountably infinite and there exists a non-negative-valued function $f(x)$ defined for all $x \in (-\infty, \infty)$ such that for any event $B \subset R(X)$, $\Pr(B) = \int_{x \in B} f(x) dx$ and $f(x) = 0$ for all $x \notin R(X)$ where $R(X)$ is the range of X (i.e. the values for which X is defined).

The pmf of a discrete random variable is replaced with the probability density function (pdf) for continuous random variables. This change in naming reflects that the probability of a single point of a continuous random variable is 0, although the probability of observing a value inside an arbitrarily small interval in $R(X)$ is not.



Figure 1.2: These four charts show examples of Bernoulli random variables using returns on the FTSE 100 and S&P 500. In the top two, a success was defined as a positive return. In the bottom two, a success was a return above -1% (weekly) or -4% (monthly).

Definition 1.16 (Probability Density Function). For a continuous random variable, the function f is called the probability density function (pdf).

Before providing some examples of pdfs, it is useful to characterize the properties that any pdf should have.

Definition 1.17 (Continuous Density Function Characterization). A function $f : \mathbb{R} \rightarrow \mathbb{R}$ is a member of the class of continuous density functions if and only if $f(x) \geq 0$ for all $x \in (-\infty, \infty)$ and $\int_{-\infty}^{\infty} f(x) dx = 1$.

There are two essential properties. First, that the function is non-negative, which follows from the axiomatic definition of probability, and second, that the function integrates to 1, so that the total probability across $R(X)$ is 1. This may seem like a limitation, but it is only a normalization since any non-negative integrable function can always be normalized to that it integrates to 1.

Example 1.12. A simple continuous random variable can be defined on $[0, 1]$ using the probability

density function

$$f(x) = 12 \left(x - \frac{1}{2} \right)^2$$

and figure 1.3 contains a plot of the pdf.

This simple pdf has peaks near 0 and 1 and a trough at 1/2. More realistic pdfs allow for values in $(-\infty, \infty)$, such as in the density of a normal random variable.

Example 1.13. The pdf of a normal random variable with parameters μ and σ^2 is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \quad (1.8)$$

$N(\mu, \sigma^2)$ is used as a shorthand notation for a random variable with this pdf. When $\mu = 0$ and $\sigma^2 = 1$, the distribution is known as a standard normal. Figure 1.3 contains a plot of the standard normal pdf along with two other parameterizations.

For large values of x (in the absolute sense), the pdf of a standard normal takes very small values, and peaks at $x = 0$ with a value of 0.3989. The shape of the normal distribution is that of a bell (and is occasionally referred to a bell curve).

A closely related function to the pdf is the cumulative distribution function, which returns the total probability of observing a value of the random variable *less* than its input.

Definition 1.18 (Cumulative Distribution Function). The cumulative distribution function (cdf) for a random variable X is defined as $F(c) = \Pr(X \leq c)$ for all $c \in (-\infty, \infty)$.

Cumulative distribution function is used for both discrete and continuous random variables.

Definition 1.19 (Discrete cdf). When X is a discrete random variable, the cdf is

$$F(x) = \sum_{s \leq x} f(s) \quad (1.9)$$

for $x \in (-\infty, \infty)$.

Example 1.14. The cdf of a Bernoulli is

$$F(x; p) = \begin{cases} 0 & \text{if } x < 0 \\ p & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}.$$

The Bernoulli cdf is simple since it only takes 3 values. The cdf of a Poisson random variable relatively simple since it is defined as sum the probability mass function for all values less than or equal to the function's argument.

Example 1.15. The cdf of a Poisson(λ)random variable is given by

$$F(x; \lambda) = \exp(-\lambda) \sum_{i=0}^{\lfloor x \rfloor} \frac{\lambda^i}{i!}, \quad x \geq 0.$$

where $\lfloor \cdot \rfloor$ returns the largest integer smaller than the input (the floor operator).

Continuous cdfs operate much like discrete cdfs, only the summation is replaced by an integral since there are a continuum of values possible for X .

Definition 1.20 (Continuous cdf). When X is a continuous random variable, the cdf is

$$F(x) = \int_{-\infty}^x f(s) ds \quad (1.10)$$

for $x \in (-\infty, \infty)$.

The integral computes the total area under the pdf starting from $-\infty$ up to x .

Example 1.16. The cdf of the random variable with pdf given by $12(x - 1/2)^2$ is

$$F(x) = 4x^3 - 6x^2 + 3x.$$

and figure 1.3 contains a plot of this cdf.

This cdf is the integral of the pdf, and checking shows that $F(0) = 0$, $F(1/2) = 1/2$ (since it is symmetric around $1/2$) and $F(1) = 1$, which must be 1 since the random variable is only defined on $[0, 1]$.⁴

Example 1.17. The cdf of a normally distributed random variable with parameters μ and σ^2 is given by

$$F(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^x \exp\left(-\frac{(s-\mu)^2}{2\sigma^2}\right) ds. \quad (1.11)$$

Figure 1.3 contains a plot of the standard normal cdf along with two other parameterizations.

In the case of a standard normal random variable, the cdf is not available in closed form, and so when computed using a computer (i.e. in Excel or MATLAB), fast, accurate numeric approximations based on polynomial expansions are used (Abramowitz and Stegun, 1964).

The cdf can be similarly derived from the pdf as long as the cdf is continuously differentiable. At points where the cdf is not continuously differentiable, the pdf is defined to take the value 0.⁴

Theorem 1.4 (Relationship between cdf and pdf). *Let $f(x)$ and $F(x)$ represent the pdf and cdf of a continuous random variable X , respectively. The density function for X can be defined as $f(x) = \frac{\partial F(x)}{\partial x}$ whenever $f(x)$ is continuous and $f(x) = 0$ elsewhere.*

Example 1.18. Taking the derivative of the cdf in the running example,

$$\begin{aligned} \frac{\partial F(x)}{\partial x} &= 12x^2 - 12x + 3 \\ &= 12\left(x^2 - x + \frac{1}{4}\right) \\ &= 12\left(x - \frac{1}{2}\right)^2. \end{aligned}$$

⁴Formally a pdf does not have to exist for a random variable, although a cdf always does. In practice, this is a technical point and distributions which have this property are rarely encountered in financial economics.

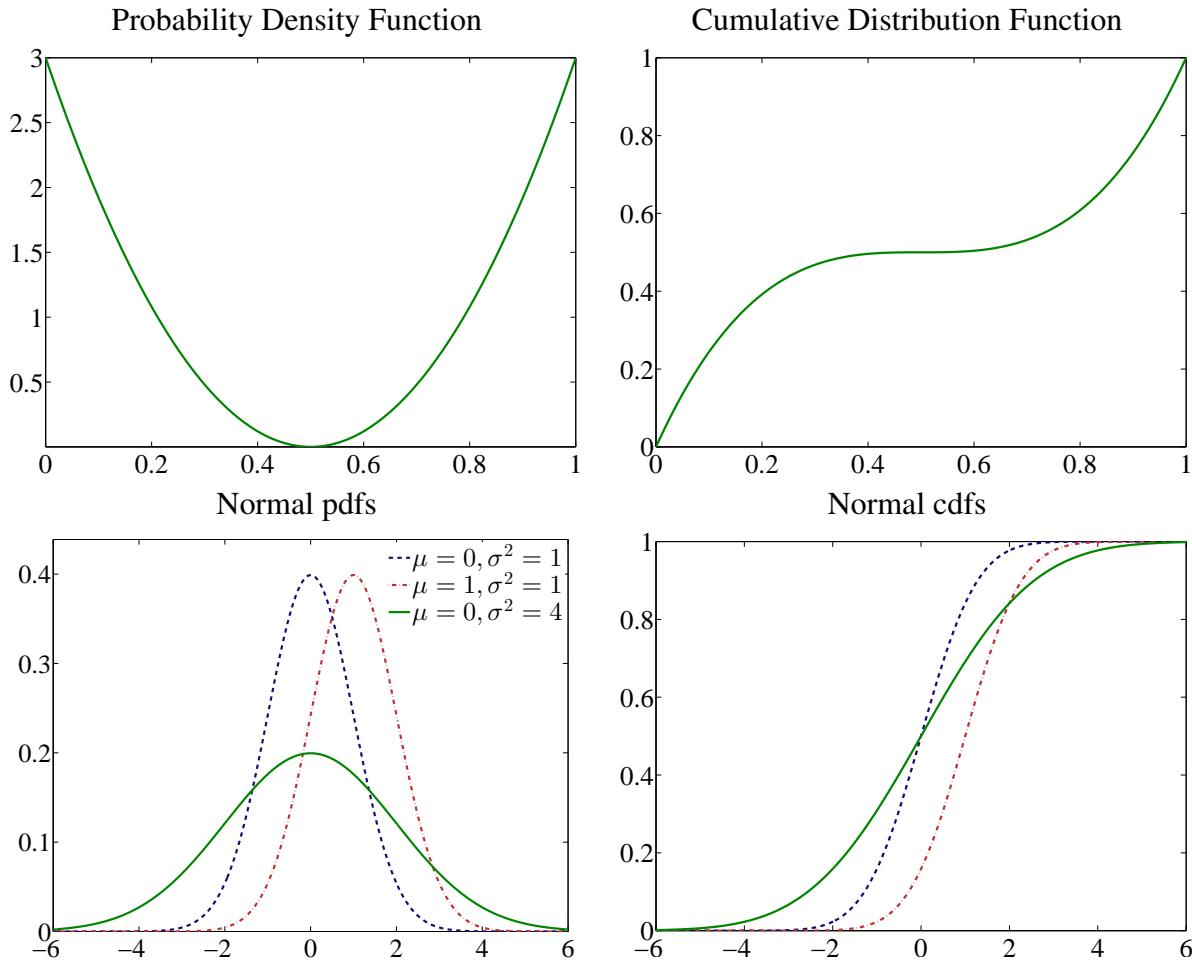


Figure 1.3: The top panels show the pdf for the density $f(x) = 12\left(x - \frac{1}{2}\right)^2$ and its associated cdf. The bottom left panel shows the probability density function for normal distributions with alternative values for μ and σ^2 . The bottom right panel shows the cdf for the same parameterizations.

1.2.2 Quantile Functions

The quantile function is closely related to the cdf – and in many important cases, the quantile function is the inverse (function) of the cdf. Before defining quantile functions, it is necessary to define a quantile.

Definition 1.21 (Quantile). Any number q satisfying $\Pr(x \leq q) = \alpha$ and $\Pr(x \geq q) = 1 - \alpha$ is known as the α -quantile of X and is denoted q_α .

A quantile is just the point on the cdf where the total probability that a random variable is smaller is α and the probability that the random variable takes a larger value is $1 - \alpha$. The definition of a quantile does not necessarily require uniqueness and non-unique quantiles are encountered when pdfs have regions of 0 probability (or equivalently cdfs are discontinuous). Quantiles are unique for random variables which have continuously differentiable cdfs. One common modification of the quantile definition is to select the *smallest* number which satisfies the two conditions to impose uniqueness of the quantile.

The function which returns the quantile is known as the quantile function.

Definition 1.22 (Quantile Function). Let X be a continuous random variable with cdf $F(x)$. The quantile function for X is defined as $G(\alpha) = q$ where $\Pr(x \leq q) = \alpha$ and $\Pr(x > q) = 1 - \alpha$. When $F(x)$ is one-to-one (and hence X is strictly continuous) then $G(\alpha) = F^{-1}(\alpha)$.

Quantile functions are generally set-valued when quantiles are not unique, although in the common case where the pdf does not contain any regions of 0 probability, the quantile function is the inverse of the cdf.

Example 1.19. The cdf of an exponential random variable is

$$F(x; \lambda) = 1 - \exp\left(-\frac{x}{\lambda}\right)$$

for $x \geq 0$ and $\lambda > 0$. Since $f(x; \lambda) > 0$ for $x > 0$, the quantile function is

$$F^{-1}(\alpha; \lambda) = -\lambda \ln(1 - \alpha).$$

The quantile function plays an important role in simulation of random variables. In particular, if $u \sim U(0, 1)$ ⁵, then $x = F^{-1}(u)$ is distributed F . For example, when u is a standard uniform ($U(0, 1)$), and $F^{-1}(\alpha)$ is the quantile function of an exponential random variable with shape parameter λ , then $x = F^{-1}(u; \lambda)$ follows an exponential(λ) distribution.

Theorem 1.5 (Probability Integral Transform). *Let U be a standard uniform random variable, $F_X(x)$ be a continuous, increasing cdf. Then $\Pr(F^{-1}(U) < x) = F_X(x)$ and so $F^{-1}(U)$ is distributed F .*

Proof. Let U be a standard uniform random variable, and for an $x \in R(X)$,

$$\Pr(U \leq F(x)) = F(x),$$

which follows from the definition of a standard uniform.

$$\begin{aligned} \Pr(U \leq F(x)) &= \Pr(F^{-1}(U) \leq F^{-1}(F(x))) \\ &= \Pr(F^{-1}(U) \leq x) \\ &= \Pr(X \leq x). \end{aligned}$$

□

The key identity is that $\Pr(F^{-1}(U) \leq x) = \Pr(X \leq x)$, which shows that the distribution of $F^{-1}(U)$ is F by definition of the cdf. The right panel of figure 1.8 shows the relationship between the cdf of a standard normal and the associated quantile function. Applying $F(X)$ produces a uniform U through the cdf and applying $F^{-1}(U)$ produces X through the quantile function.

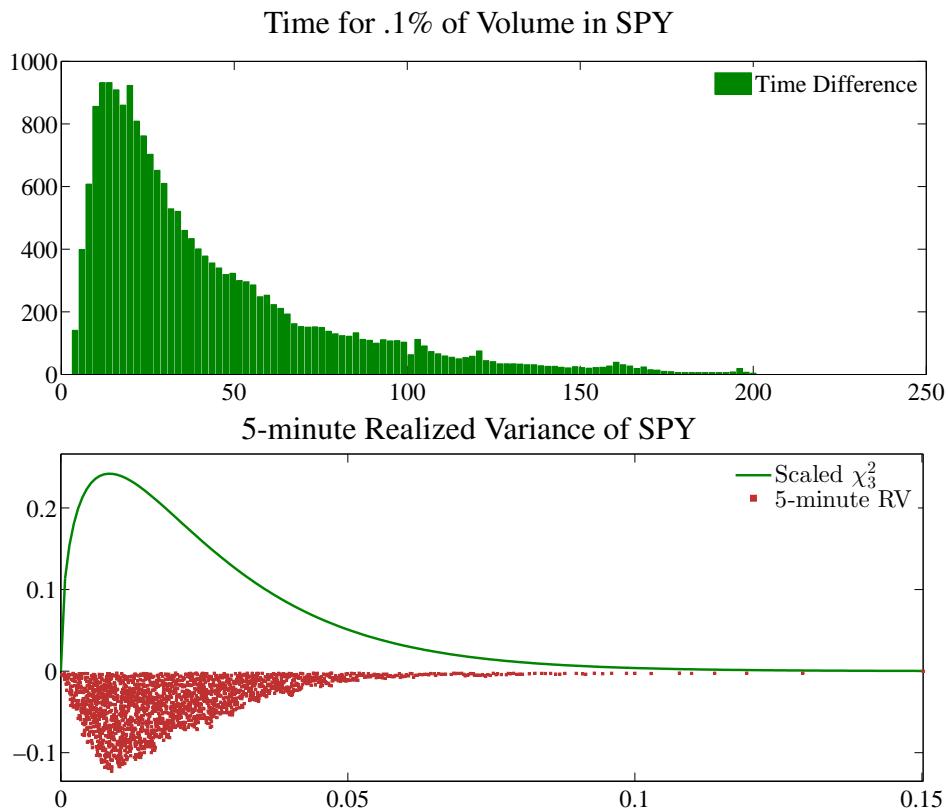


Figure 1.4: The left panel shows a histogram of the elapsed time in seconds required for .1% of the daily volume being traded to occur for SPY on May 31, 2012. The right panel shows both the fitted scaled χ^2 distribution and the raw data (mirrored below) for 5-minute "realized variance" estimates for SPY on May 31, 2012.

1.2.3 Common Univariate Distributions

Discrete

1.2.3.1 Bernoulli

A Bernoulli random variable is a discrete random variable which takes one of two values, 0 or 1. It is often used to model success or failure, where success is loosely defined. For example, a success may be the event that a trade was profitable net of costs, or the event that stock market volatility as measured by VIX was greater than 40%. The Bernoulli distribution depends on a single parameter p which determines the probability of success.

Parameters

$$p \in [0, 1]$$

⁵The mathematical notation \sim is read “distributed as”. For example, $x \sim U(0, 1)$ indicates that x is distributed as a standard uniform random variable.

Support

$$x \in \{0, 1\}$$

Probability Mass Function

$$f(x; p) = p^x (1 - p)^{1-x}, p \geq 0$$

Moments

Mean	p
Variance	$p(1 - p)$

1.2.3.2 Poisson

A Poisson random variable is a discrete random variable taking values in $\{0, 1, \dots\}$. The Poisson depends on a single parameter λ (known as the intensity). Poisson random variables are often used to model counts of events during some interval, for example the number of trades executed over a 5-minute window.

Parameters

$$\lambda \geq 0$$

Support

$$x \in \{0, 1, \dots\}$$

Probability Mass Function

$$f(x; \lambda) = \frac{\lambda^x}{x!} \exp(-\lambda)$$

Moments

Mean	λ
Variance	λ

Continuous

1.2.3.3 Normal (Gaussian)

The normal is the most important univariate distribution in financial economics. It is the familiar “bell-shaped” distribution, and is used heavily in hypothesis testing and in modeling (net) asset returns (e.g. $r_t = \ln P_t - \ln P_{t-1}$ or $r_t = \frac{P_t - P_{t-1}}{P_{t-1}}$ where P_t is the price of the asset in period t).

Parameters

$$\mu \in (-\infty, \infty), \sigma^2 \geq 0$$

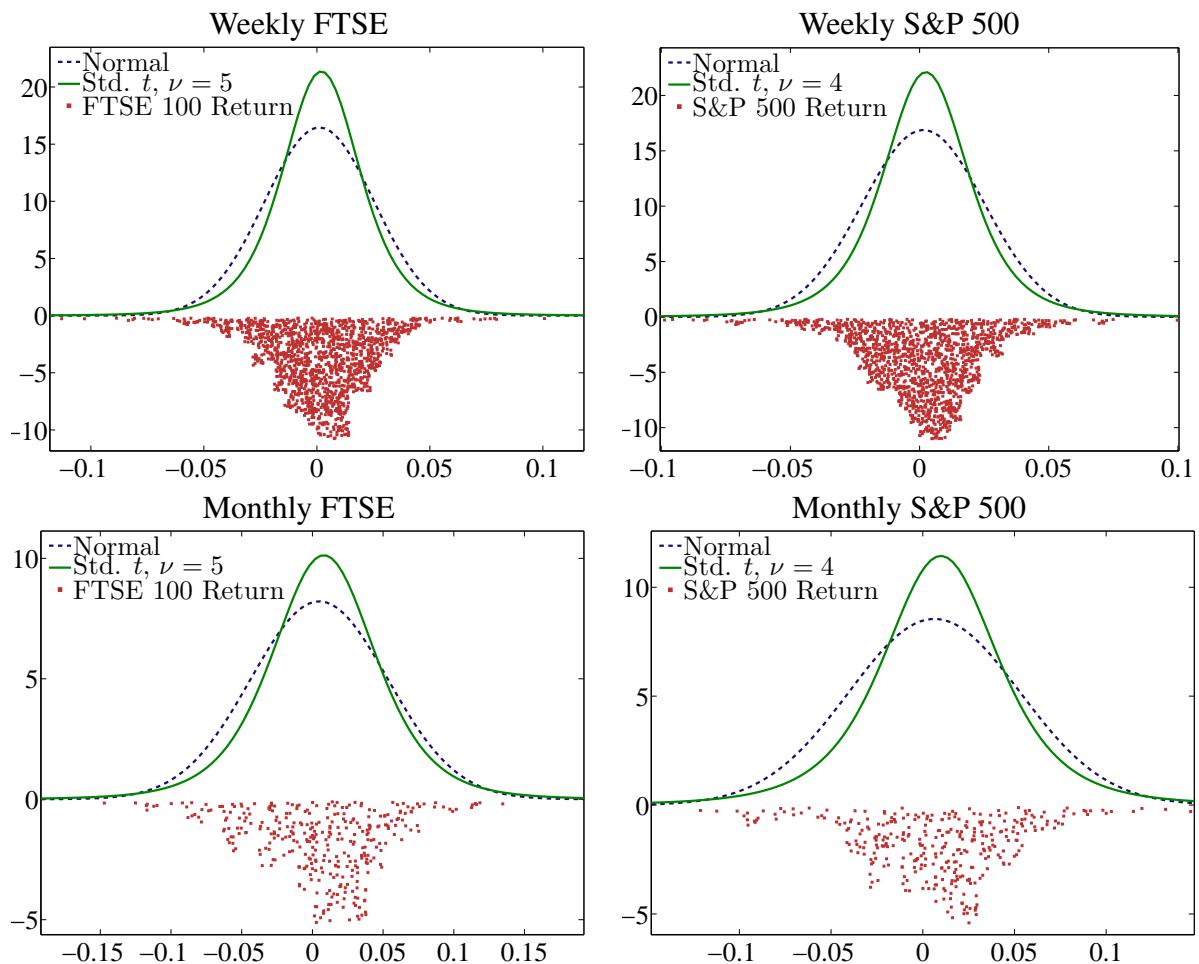


Figure 1.5: Weekly and monthly densities for the FTSE 100 and S&P 500. All panels plot the pdf of a normal and a standardized Student's t using parameters estimated with maximum likelihood estimation (See Chapter 1). The points below 0 on the y-axis show the actual returns observed during this period.

Support

$$x \in (-\infty, \infty)$$

Probability Density Function

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Cumulative Distribution Function

$$F(x; \mu, \sigma^2) = \frac{1}{2} + \frac{1}{2} \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right) \text{ where } \operatorname{erf} \text{ is the error function.}^6$$

Moments

Mean	μ
Variance	σ^2
Median	μ
Skewness	0
Kurtosis	3

Notes

The normal with mean μ and variance σ^2 is written $N(\mu, \sigma^2)$. A normally distributed random variable with $\mu = 0$ and $\sigma^2 = 1$ is known as a standard normal. Figure 1.5 shows the fit normal distribution to the FTSE 100 and S&P 500 using both weekly and monthly returns for the period 1984–2012. Below each figure is a plot of the raw data.

1.2.3.4 Log-Normal

Log-normal random variables are closely related to normals. If X is log-normal, then $Y = \ln(X)$ is normal. Like the normal, the log-normal family depends on two parameters, μ and σ^2 , although unlike the normal these parameters do not correspond to the mean and variance. Log-normal random variables are commonly used to model gross returns, P_{t+1}/P_t (although it is often simpler to model $r_t = \ln P_t - \ln P_{t-1} = \ln(P_t/P_{t-1})$ which is normally distributed).

Parameters

$$\mu \in (-\infty, \infty), \sigma^2 \geq 0$$

Support

$$x \in (0, \infty)$$

⁶The error function does not have a closed form and is defined

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-s^2) ds.$$

Probability Density Function

$$f(x; \mu, \sigma^2) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$$

Cumulative Distribution Function

Since $Y = \ln(X) \sim N(\mu, \sigma^2)$, the cdf is the same as the normal only using $\ln x$ in place of x .

Moments

Mean	$\exp\left(\mu + \frac{\sigma^2}{2}\right)$
Median	$\exp(\mu)$
Variance	$\{\exp(\sigma^2) - 1\} \exp(2\mu + \sigma^2)$

1.2.3.5 χ^2 (Chi-square)

χ_v^2 random variables depend on a single parameter v known as the degree-of-freedom. They are commonly encountered when testing hypotheses, although they are also used to model continuous variables which are non-negative such as conditional variances. χ_v^2 random variables are closely related to standard normal random variables and are defined as the sum of v independent standard normal random variables which have been squared. Suppose Z_1, \dots, Z_v are standard normally distributed and independent, then $x = \sum_{i=1}^v z_i^2$ follows a χ_v^2 .⁷

Parameters

$$v \in [0, \infty)$$

Support

$$x \in [0, \infty)$$

Probability Density Function

$$f(x; v) = \frac{1}{2^{\frac{v}{2}} \Gamma(\frac{v}{2})} x^{\frac{v-2}{2}} \exp\left(-\frac{x}{2}\right), v \in \{1, 2, \dots\} \text{ where } \Gamma(a) \text{ is the Gamma function.}^8$$

Cumulative Distribution Function

$$F(x; v) = \frac{1}{\Gamma(\frac{v}{2})} \gamma\left(\frac{v}{2}, \frac{x}{2}\right) \text{ where } \gamma(a, b) \text{ is the lower incomplete gamma function.}$$

Moments

Mean	v
Variance	$2v$

⁷ v does not need to be an integer,

⁸The χ_v^2 is related to the gamma distribution which has pdf $f(x; \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} \exp(-x/\beta)$ by setting $\alpha = v/2$ and $\beta = 2$.

Notes

Figure 1.4 shows a χ^2 pdf which was used to fit some simple estimators of the 5-minute variance of the S&P 500 from May 31, 2012. These were computed by summing and squaring 1-minute returns within a 5-minute interval (all using log prices). 5-minute variance estimators are important in high-frequency trading and other (slower) algorithmic trading.

1.2.3.6 Student's t and standardized Student's t

Student's t random variables are also commonly encountered in hypothesis testing and, like χ^2 random variables, are closely related to standard normals. Student's t random variables depend on a single parameter, v , and can be constructed from two other independent random variables. If Z a standard normal, W a χ^2_v and $Z \perp\!\!\!\perp W$, then $x = z / \sqrt{w/v}$ follows a Student's t distribution. Student's t are similar to normals except that they are heavier tailed, although as $v \rightarrow \infty$ a Student's t converges to a standard normal.

Support

$$x \in (-\infty, \infty)$$

Probability Density Function

$$f(x; v) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{v}\pi\Gamma(\frac{v}{2})} \left(1 + \frac{x^2}{v}\right)^{-\frac{v+1}{2}} \text{ where } \Gamma(a) \text{ is the Gamma function.}$$

Moments

Mean	$0, v > 1$
Median	0
Variance	$\frac{v}{v-2}, v > 2$
Skewness	0, $v > 3$
Kurtosis	$3\frac{(v-2)}{v-4}, v > 4$

Notes

When $v = 1$, a Student's t is known as a Cauchy random variable. Cauchy random variables are so heavy-tailed that even the mean does not exist.

The standardized Student's t extends the usual Student's t in two directions. First, it removes the variance's dependence on v so that the scale of the random variable can be established separately from the degree of freedom parameter. Second, it explicitly adds location and scale parameters so that if Y is a Student's t random variable with degree of freedom v , then

$$x = \mu + \sigma \frac{\sqrt{v-2}}{\sqrt{v}} y$$

follows a standardized Student's t distribution ($v > 2$ is required). The standardized Student's t is commonly used to model heavy-tailed return distributions such as stock market indices.

Figure 1.5 shows the fit (using maximum likelihood) standardized t distribution to the FTSE 100 and S&P 500 using both weekly and monthly returns from the period 1984–2012. The typical degree of freedom parameter was around 4, indicating that (unconditional) distributions are heavy-tailed with a large kurtosis.

1.2.3.7 Uniform

The continuous uniform is commonly encountered in certain test statistics, especially those testing whether assumed densities are appropriate for a particular series. Uniform random variables, when combined with quantile functions, are also useful for simulating random variables.

Parameters

a, b the end points of the interval, where $a < b$

Support

$x \in [a, b]$

Probability Density Function

$$f(x) = \frac{1}{b-a}$$

Cumulative Distribution Function

$$F(x) = \frac{x-a}{b-a} \text{ for } a \leq x \leq b, F(x) = 0 \text{ for } x < a \text{ and } F(x) = 1 \text{ for } x > b$$

Moments

Mean	$\frac{b-a}{2}$
Median	$\frac{b-a}{2}$
Variance	$\frac{(b-a)^2}{12}$
Skewness	0
Kurtosis	$\frac{9}{5}$

Notes

A standard uniform has $a = 0$ and $b = 1$. When $x \sim F$, then $F(x) \sim U(0, 1)$

1.3 Multivariate Random Variables

While univariate random variables are very important in financial economics, most applications require the use multivariate random variables. Multivariate random variables allow the relationship between two or more random quantities to be modeled and studied. For example, the joint distribution of equity and bond returns is important for many investors.

Throughout this section, the multivariate random variable is assumed to have n components,

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

which are arranged into a column vector. The definition of a multivariate random variable is virtually identical to that of a univariate random variable, only mapping $\omega \in \Omega$ to the n -dimensional space \mathbb{R}^n .

Definition 1.23 (Multivariate Random Variable). Let (Ω, \mathcal{F}, P) be a probability space. If $X : \Omega \rightarrow \mathbb{R}^n$ is a real-valued vector function having its domain the elements of Ω , then $X : \Omega \rightarrow \mathbb{R}^n$ is called a (multivariate) n -dimensional random variable.

Multivariate random variables, like univariate random variables, are technically functions of events in the underlying probability space $X(\omega)$, although the function argument ω (the event) is usually suppressed.

Multivariate random variables can be either discrete or continuous. Discrete multivariate random variables are fairly uncommon in financial economics and so the remainder of the chapter focuses exclusively on the continuous case. The characterization of what makes a multivariate random variable continuous is also virtually identical to that in the univariate case.

Definition 1.24 (Continuous Multivariate Random Variable). A multivariate random variable is said to be continuous if its range is uncountably infinite and if there exists a non-negative valued function $f(x_1, \dots, x_n)$ defined for all $(x_1, \dots, x_n) \in \mathbb{R}^n$ such that for any event $B \subset R(X)$,

$$\Pr(B) = \int \dots \int_{\{x_1, \dots, x_n\} \in B} f(x_1, \dots, x_n) dx_1 \dots dx_n \quad (1.12)$$

and $f(x_1, \dots, x_n) = 0$ for all $(x_1, \dots, x_n) \notin R(X)$.

Multivariate random variables, at least when continuous, are often described by their probability density function.

Definition 1.25 (Continuous Density Function Characterization). A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a member of the class of multivariate continuous density functions if and only if $f(x_1, \dots, x_n) \geq 0$ for all $x \in \mathbb{R}^n$ and

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, \dots, x_n) dx_1 \dots dx_n = 1. \quad (1.13)$$

Definition 1.26 (Multivariate Probability Density Function). The function $f(x_1, \dots, x_n)$ is called a multivariate probability function (pdf).

A multivariate density, like a univariate density, is a function which is everywhere non-negative and which integrates to unity. Figure 1.7 shows the fit joint probability density function to weekly returns on the FTSE 100 and S&P 500 (assuming that returns are normally distributed). Two views are presented – one shows the 3-dimensional plot of the pdf and the other shows the iso-probability contours of the pdf. The figure also contains a scatter plot of the raw weekly data for comparison. All parameters were estimated using maximum likelihood.

Example 1.20. Suppose X is a bivariate random variable, then the function $f(x_1, x_2) = \frac{3}{2}(x_1^2 + x_2^2)$ defined on $[0, 1] \times [0, 1]$ is a valid probability density function.

Example 1.21. Suppose X is a bivariate standard normal random variable. Then the probability density function of X is

$$f(x_1, x_2) = \frac{1}{2\pi} \exp\left(-\frac{x_1^2 + x_2^2}{2}\right).$$

The multivariate cumulative distribution function is virtually identical to that in the univariate case, and measure the total probability between $-\infty$ (for each element of X) and some point.

Definition 1.27 (Multivariate Cumulative Distribution Function). The joint cumulative distribution function of an n -dimensional random variable X is defined by

$$F(x_1, \dots, x_n) = \Pr(X_i \leq x_i, i = 1, \dots, n)$$

for all $(x_1, \dots, x_n) \in \mathbb{R}^n$, and is given by

$$F(x_1, \dots, x_n) = \int_{-\infty}^{x_n} \dots \int_{-\infty}^{x_1} f(s_1, \dots, s_n) ds_1 \dots ds_n. \quad (1.14)$$

Example 1.22. Suppose X is a bivariate random variable with probability density function

$$f(x_1, x_2) = \frac{3}{2}(x_1^2 + x_2^2)$$

defined on $[0, 1] \times [0, 1]$. Then the associated cdf is

$$F(x_1, x_2) = \frac{x_1^3 x_2 + x_1 x_2^3}{2}.$$

Figure 1.6 shows the joint cdf of the density in the previous example. As was the case for univariate random variables, the probability density function can be determined by differentiating the cumulative distribution function with respect to each component.

Theorem 1.6 (Relationship between cdf and pdf). *Let $f(x_1, \dots, x_n)$ and $F(x_1, \dots, x_n)$ represent the pdf and cdf of an n -dimensional continuous random variable X , respectively. The density function for X can be defined as $f(x_1, \dots, x_n) = \frac{\partial^n F(\mathbf{x})}{\partial x_1 \partial x_2 \dots \partial x_n}$ whenever $f(x_1, \dots, x_n)$ is continuous and $f(x_1, \dots, x_n) = 0$ elsewhere.*

Example 1.23. Suppose X is a bivariate random variable with cumulative distribution function $F(x_1, x_2) = \frac{x_1^3 x_2 + x_1 x_2^3}{2}$. The probability density function can be determined using

$$\begin{aligned} f(x_1, x_2) &= \frac{\partial^2 F(x_1, x_2)}{\partial x_1 \partial x_2} \\ &= \frac{1}{2} \frac{\partial(3x_1^2 x_2 + x_1 x_2^3)}{\partial x_2} \\ &= \frac{3}{2} (x_1^2 + x_2^2). \end{aligned}$$

1.3.1 Marginal Densities and Distributions

The marginal distribution is the first concept unique to multivariate random variables. Marginal densities and distribution functions summarize the information in a subset, usually a single component, of X by averaging over all possible values of the components of X which are not being marginalized. This involves integrating out the variables which are not of interest. First, consider the bivariate case.

Definition 1.28 (Bivariate Marginal Probability Density Function). Let X be a bivariate random variable comprised of X_1 and X_2 . The marginal distribution of X_1 is given by

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2. \quad (1.15)$$

The marginal density of X_1 is a density function where X_2 has been integrated out. This integration is simply a form of averaging – varying x_2 according to the probability associated with each value of x_2 – and so the marginal is only a function of x_1 . Both probability density functions and cumulative distribution functions have marginal versions.

Example 1.24. Suppose X is a bivariate random variable with probability density function

$$f(x_1, x_2) = \frac{3}{2} (x_1^2 + x_2^2)$$

and is defined on $[0, 1] \times [0, 1]$. The marginal probability density function for X_1 is

$$f_1(x_1) = \frac{3}{2} \left(x_1^2 + \frac{1}{3} \right),$$

and by symmetry the marginal probability density function of X_2 is

$$f_2(x_2) = \frac{3}{2} \left(x_2^2 + \frac{1}{3} \right).$$

Example 1.25. Suppose X is a bivariate random variable with probability density function $f(x_1, x_2) = 6(x_1 x_2^2)$ and is defined on $[0, 1] \times [0, 1]$. The marginal probability density functions for X_1 and X_2 are

$$f_1(x_1) = 2x_1 \text{ and } f_2(x_2) = 3x_2^2.$$

Example 1.26. Suppose X is bivariate normal with parameters $\mu = [\mu_1 \mu_2]'$ and

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix},$$

then the marginal pdf of X_1 is $N(\mu_1, \sigma_1^2)$, and the marginal pdf of X_2 is $N(\mu_2, \sigma_2^2)$.

Figure 1.7 shows the fit marginal distributions to weekly returns on the FTSE 100 and S&P 500 assuming that returns are normally distributed. Marginal pdfs can be transformed into marginal cdfs through integration.

Definition 1.29 (Bivariate Marginal Cumulative Distribution Function). The cumulative marginal distribution function of X_1 in bivariate random variable X is defined by

$$F_1(x_1) = \Pr(X_1 \leq x_1)$$

for all $x_1 \in \mathbb{R}$, and is given by

$$F_1(x_1) = \int_{-\infty}^{x_1} f_1(s_1) ds_1.$$

The general j -dimensional marginal distribution partitions the n -dimensional random variable X into two blocks, and constructs the marginal distribution for the first j by integrating out (averaging over) the remaining $n - j$ components of X . In the definition, both X_1 and X_2 are vectors.

Definition 1.30 (Marginal Probability Density Function). Let X be a n -dimensional random variable and partition the first j ($1 \leq j < n$) elements of X into X_1 , and the remainder into X_2 so that $X = [X'_1 X'_2]'$. The marginal probability density function for X_1 is given by

$$f_{1,\dots,j}(x_1, \dots, x_j) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, \dots, x_n) dx_{j+1} \dots dx_n. \quad (1.16)$$

The marginal cumulative distribution function is related to the marginal probability density function in the same manner as the joint probability density function is related to the cumulative distribution function. It also has the same interpretation.

Definition 1.31 (Marginal Cumulative Distribution Function). Let X be a n -dimensional random variable and partition the first j ($1 \leq j < n$) elements of X into X_1 , and the remainder into X_2 so that $X = [X'_1 X'_2]'$. The marginal cumulative distribution function for X_1 is given by

$$F_{1,\dots,j}(x_1, \dots, x_j) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_j} f_{1,\dots,j}(s_1, \dots, s_j) ds_1 \dots ds_j. \quad (1.17)$$

1.3.2 Conditional Distributions

Marginal distributions provide the tools needed to model the distribution of a subset of the components of a random variable while averaging over the other components. Conditional densities and distributions, on the other hand, consider a subset of the components a random variable conditional on observing a specific value for the remaining components. In practice, the vast majority of modeling makes use of conditioning information where the interest is in understanding the distribution of a random variable conditional on the observed values of some other random variables. For example, consider the problem of modeling the expected return of an individual stock. Balance sheet information such as the book value of assets, earnings and return on equity are all available, and can be conditioned on to model the conditional distribution of the stock's return.

First, consider the bivariate case.

Definition 1.32 (Bivariate Conditional Probability Density Function). Let X be a bivariate random variable comprised of X_1 and X_2 . The conditional probability density function for X_1 given that $X_2 \in B$ where B is an event where $\Pr(X_2 \in B) > 0$ is

$$f(x_1 | X_2 \in B) = \frac{\int_B f(x_1, x_2) dx_2}{\int_B f_2(x_2) dx_2}. \quad (1.18)$$

When B is an elementary event (e.g. single point), so that $\Pr(X_2 = x_2) = 0$ and $f_2(x_2) > 0$, then

$$f(x_1|X_2 = x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}. \quad (1.19)$$

Conditional density functions differ slightly depending on whether the conditioning variable is restricted to a set or a point. When the conditioning variable is specified to be a set where $\Pr(X_2 \in B) > 0$, then the conditional density is the joint probability of X_1 and $X_2 \in B$ divided by the marginal probability of $X_2 \in B$. When the conditioning variable is restricted to a point, the conditional density is the ratio of the joint pdf to the margin pdf of X_2 .

Example 1.27. Suppose X is a bivariate random variable with probability density function

$$f(x_1, x_2) = \frac{3}{2} (x_1^2 + x_2^2)$$

and is defined on $[0, 1] \times [0, 1]$. The conditional probability of X_1 given $X_2 \in [\frac{1}{2}, 1]$

$$f\left(x_1|X_2 \in \left[\frac{1}{2}, 1\right]\right) = \frac{1}{11} (12x_1^2 + 7),$$

the conditional probability density function of X_1 given $X_2 \in [0, \frac{1}{2}]$ is

$$f\left(x_1|X_2 \in \left[0, \frac{1}{2}\right]\right) = \frac{1}{5} (12x_1^2 + 1),$$

and the conditional probability density function of X_1 given $X_2 = x_2$ is

$$f(x_1|X_2 = x_2) = \frac{x_1^2 + x_2^2}{x_2^2 + 1}.$$

Figure 1.6 shows the joint pdf along with both types of conditional densities. The upper left panel shows that conditional density for $X_2 \in [0.25, 0.5]$. The highlighted region contains the components of the joint pdf which are averaged to produce the conditional density. The lower left also shows the pdf but also shows three (non-normalized) conditional densities of the form $f(x_1|x_2)$. The lower right pane shows these three densities correctly normalized.

The previous example shows that, in general, the conditional probability density function differs as the region used changes.

Example 1.28. Suppose X is bivariate normal with mean $\mu = [\mu_1 \mu_2]'$ and covariance

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix},$$

then the conditional distribution of X_1 given $X_2 = x_2$ is $N\left(\mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(x_2 - \mu_2), \sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2}\right)$.

Marginal distributions and conditional distributions are related in a number of ways. One obvious way is that $f(x_1|X_2 \in R(X_2)) = f_1(x_1)$ – that is, the conditional probability of X_1 given that X_2 is in its range is the marginal pdf of X_1 . This holds since integrating over all values of x_2 is essentially not conditioning on anything (which is known as the unconditional, and a marginal density could, in principle, be called the unconditional density since it averages across all values of the other variable).

The general definition allows for an n -dimensional random vector where the conditioning variable has a dimension between 1 and $j < n$.

Definition 1.33 (Conditional Probability Density Function). Let $f(x_1, \dots, x_n)$ be the joint density function for an n -dimensional random variable $X = [X_1 \dots X_n]'$ and partition the first j ($1 \leq j < n$) elements of X into X_1 , and the remainder into X_2 so that $X = [X'_1 X'_2]'$. The conditional probability density function for X_1 given that $X_2 \in B$ is given by

$$f(x_1, \dots, x_j | X_2 \in B) = \frac{\int_{(x_{j+1}, \dots, x_n) \in B} f(x_1, \dots, x_n) dx_n \dots dx_{j+1}}{\int_{(x_{j+1}, \dots, x_n) \in B} f_{j+1, \dots, n}(x_{j+1}, \dots, x_n) dx_n \dots dx_{j+1}}, \quad (1.20)$$

and when B is an elementary event (denoted \mathbf{x}_2) and if $f_{j+1, \dots, n}(\mathbf{x}_2) > 0$,

$$f(x_1, \dots, x_j | X_2 = \mathbf{x}_2) = \frac{f(x_1, \dots, x_j, \mathbf{x}_2)}{f_{j+1, \dots, n}(\mathbf{x}_2)} \quad (1.21)$$

In general the simplified notation $f(x_1, \dots, x_j | \mathbf{x}_2)$ will be used to represent $f(x_1, \dots, x_j | X_2 = \mathbf{x}_2)$.

1.3.3 Independence

A special relationship exists between the joint probability density function and the marginal density functions when random variables are independent— the joint must be the product of each marginal.

Theorem 1.7 (Independence of Random Variables). *The random variables X_1, \dots, X_n with joint density function $f(x_1, \dots, x_n)$ are independent if and only if*

$$f(x_1, \dots, x_n) = \prod_{i=1}^n f_i(x_i) \quad (1.22)$$

where $f_i(x_i)$ is the marginal distribution of X_i .

The intuition behind this result follows from the fact that when the components of a random variable are independent, any change in one component has no information for the others. In other words, both marginals and conditionals must be the same.

Example 1.29. Let X be a bivariate random variable with probability density function $f(x_1, x_2) = x_1 x_2$ on $[0, 1] \times [0, 1]$, then X_1 and X_2 are independent. This can be verified since

$$f_1(x_1) = x_1 \text{ and } f_2(x_2) = x_2$$

so that the joint is the product of the two marginal densities.

Independence is a very strong concept, and it carries over from random variables to functions of random variables as long as each function involves only one random variable.⁹

Theorem 1.8 (Independence of Functions of Independent Random Variables). *Let X_1 and X_2 be independent random variables and define $y_1 = Y_1(x_1)$ and $y_2 = Y_2(x_2)$, then the random variables Y_1 and Y_2 are independent.*

⁹This can be generalized to the full multivariate case where X is an n -dimensional random variable where the first j components are independent from the last $n - j$ components defining $y_1 = Y_1(x_1, \dots, x_j)$ and $y_2 = Y_2(x_{j+1}, \dots, x_n)$.

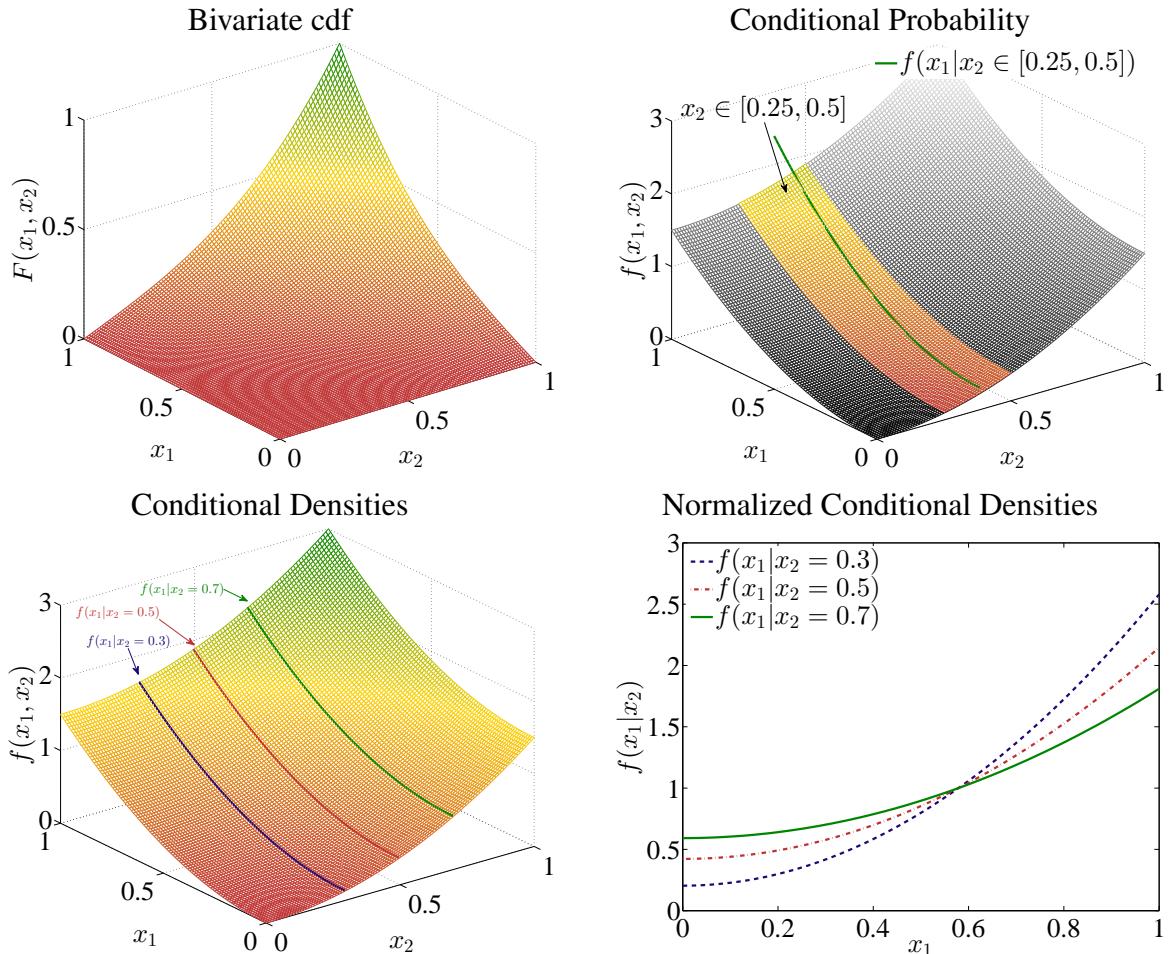


Figure 1.6: These four panels show four views of a distribution defined on $[0, 1] \times [0, 1]$. The upper left panel shows the joint cdf. The upper right shows the pdf along with the portion of the pdf used to construct a conditional distribution $f(x_1 | x_2 \in [0.25, 0.5])$. The line shows the actual correctly scaled conditional distribution which is only a function of x_1 plotted at $E[X_2 | X_2 \in [0.25, 0.5]]$. The lower left panel also shows the pdf along with three non-normalized conditional densities. The bottom right panel shows the correctly normalized conditional densities.

Independence is often combined with an assumption that the marginal distribution is the same to simplify the analysis of collections of random data.

Definition 1.34 (Independent, Identically Distributed). Let $\{X_i\}$ be a sequence of random variables. If the marginal distribution for X_i is the same for all i and $X_i \perp\!\!\!\perp X_j$ for all $i \neq j$, then $\{X_i\}$ is said to be an independent, identically distributed (i.i.d.) sequence.

1.3.4 Bayes Rule

Bayes rule is used both in financial economics and econometrics. In financial economics, it is often used to model agents learning, and in econometrics it is used to make inference about unknown parameters given observed data (a branch known as Bayesian econometrics). Bayes rule follows directly from the definition of a conditional density so that the joint can be factored into a conditional and a marginal. Suppose X is a bivariate random variable, then

$$\begin{aligned} f(x_1, x_2) &= f(x_1|x_2)f_2(x_2) \\ &= f(x_2|x_1)f_1(x_2). \end{aligned}$$

The joint can be factored two ways, and equating the two factorizations results in Bayes rule.

Definition 1.35 (Bivariate Bayes Rule). Let X by a bivariate random variable with components X_1 and X_2 , then

$$f(x_1|x_2) = \frac{f(x_2|x_1)f_1(x_1)}{f_2(x_2)} \quad (1.23)$$

Bayes rule states that the probability of observing X_1 given a value of X_2 is equal to the joint probability of the two random variables divided by the marginal probability of observing X_2 . Bayes rule is normally applied where there is a belief about X_1 ($f_1(x_1)$, called a *prior*), and the conditional distribution of X_1 given X_2 is a known density ($f(x_2|x_1)$, called the *likelihood*), which combine to form a belief about X_1 ($f(x_1|x_2)$, called the *posterior*). The marginal density of X_2 is not important when using Bayes rule since the numerator is still proportional to the conditional density of X_1 given X_2 since $f_2(x_2)$ is a number, and so it is common to express the non-normalized posterior as

$$f(x_1|x_2) \propto f(x_2|x_1)f_1(x_1),$$

where \propto is read “is proportional to”.

Example 1.30. Suppose interest lies in the probability a firm does bankrupt which can be modeled as a Bernoulli distribution. The parameter p is unknown but, given a value of p , the likelihood that a firm goes bankrupt is

$$f(x|p) = p^x(1-p)^{1-x}.$$

While p is known, a prior for the bankruptcy rate can be specified. Suppose the prior for p follows a Beta(α, β) distribution which has pdf

$$f(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)}$$

where $B(a, b)$ is Beta function that acts as a normalizing constant.¹⁰ The Beta distribution has support on $[0, 1]$ and nests the standard uniform as a special case when $\alpha = \beta = 1$. The expected value of a random variable with a Beta(α, β) is $\frac{\alpha}{\alpha+\beta}$ and the variance is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ where $\alpha > 0$ and $\beta > 0$.

Using Bayes rule,

$$\begin{aligned} f(p|x) &\propto p^x (1-p)^{1-x} \times \frac{p^{\alpha-1} (1-p)^{\beta-1}}{B(\alpha, \beta)} \\ &= \frac{p^{\alpha-1+x} (1-p)^{\beta-x}}{B(\alpha, \beta)}. \end{aligned}$$

Note that this isn't a density since it has the wrong normalizing constant. However, the component of the density which contains p is $p^{(\alpha-x)-1} (1-p)^{(\beta-x+1)-1}$ (known as the *kernel*) is the same as in the Beta distribution, only with different parameters. Thus the posterior, $f(p|x)$ is Beta($\alpha+x, \beta-x+1$). Since the posterior is the same as the prior, it could be combined with another observation (and the Bernoulli likelihood) to produce an updated posterior. When a Bayesian problem has this property, the prior density said to be conjugate to the likelihood.

Example 1.31. Suppose M is a random variable representing the score on the midterm, and interest lies in the final course grade, C . The prior for C is normal with mean μ and variance σ^2 , and that the distribution of M given C is also conditionally normal with mean C and variance τ^2 . Bayes rule can be used to make inference on the final course grade given the midterm grade.

$$\begin{aligned} f(c|m) &\propto f(m|c) f_C(c) \\ &\propto \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(m-c)^2}{2\tau^2}\right) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(c-\mu)^2}{2\sigma^2}\right) \\ &= K \exp\left(-\frac{1}{2} \left\{ \frac{(m-c)^2}{\tau^2} + \frac{(c-\mu)^2}{\sigma^2} \right\}\right) \\ &= K \exp\left(-\frac{1}{2} \left\{ \frac{c^2}{\tau^2} + \frac{c^2}{\sigma^2} - \frac{2cm}{\tau^2} - \frac{2c\mu}{\sigma^2} + \frac{m^2}{\tau^2} + \frac{\mu^2}{\sigma^2} \right\}\right) \\ &= K \exp\left(-\frac{1}{2} \left\{ c^2 \left(\frac{1}{\tau^2} + \frac{1}{\sigma^2} \right) - 2c \left(\frac{m}{\tau^2} + \frac{\mu}{\sigma^2} \right) + \left(\frac{m^2}{\tau^2} + \frac{\mu^2}{\sigma^2} \right) \right\}\right) \end{aligned}$$

This (non-normalized) density can be shown to have the kernel of a normal by completing the square,¹¹

¹⁰The beta function can only be given as an indefinite integral,

$$B(a, b) = \int_0^1 s^{a-1} (1-s)^{b-1} ds.$$

¹¹Suppose a quadratic in x has the form $ax^2 + bx + c$. Then

$$ax^2 + bx + c = a(x-d)^2 + e$$

where $d = b/(2a)$ and $e = c - b^2/(4a)$.

$$f(c|m) \propto \exp \left(-\frac{1}{2 \left(\frac{1}{\tau^2} + \frac{1}{\sigma^2} \right)^{-1}} \left(c - \frac{\left(\frac{m}{\tau^2} + \frac{\mu}{\sigma^2} \right)}{\left(\frac{1}{\tau^2} + \frac{1}{\sigma^2} \right)} \right)^2 \right).$$

This is the kernel of a normal density with mean

$$\frac{\left(\frac{m}{\tau^2} + \frac{\mu}{\sigma^2} \right)}{\left(\frac{1}{\tau^2} + \frac{1}{\sigma^2} \right)},$$

and variance

$$\left(\frac{1}{\tau^2} + \frac{1}{\sigma^2} \right)^{-1}.$$

The mean is a weighted average of the prior mean, μ and the midterm score, m , where the weights are determined by the inverse variance of the prior and conditional distributions. Since the weights are proportional to the inverse of the variance, a small variance leads to a relatively large weight. If $\tau^2 = \sigma^2$, then the posterior mean is the average of the prior mean and the midterm score. The variance of the posterior depends on the uncertainty in the prior (σ^2) and the uncertainty in the data (τ^2). The posterior variance is always less than the smaller of σ^2 and τ^2 . Like the Bernoulli-Beta combination in the previous problem, the normal distribution is a conjugate prior when the conditional density is normal.

1.3.5 Common Multivariate Distributions

1.3.5.1 Multivariate Normal

Like the univariate normal, the multivariate normal depends on 2 parameters, μ and n by 1 vector of means and Σ an n by n positive semi-definite covariance matrix. The multivariate normal is closed to both to marginalization and conditioning – in other words, if X is multivariate normal, then all marginal distributions of X are normal, and so are all conditional distributions of X_1 given X_2 for any partitioning.

Parameters

$\mu \in \mathbb{R}^n$, Σ a positive semi-definite matrix

Support

$\mathbf{x} \in \mathbb{R}^n$

Probability Density Function

$$f(\mathbf{x}; \mu, \Sigma) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp \left(-\frac{1}{2} (\mathbf{x} - \mu)' \Sigma^{-1} (\mathbf{x} - \mu) \right)$$

Cumulative Distribution Function

Can be expressed as a series of n univariate normal cdfs using repeated conditioning.

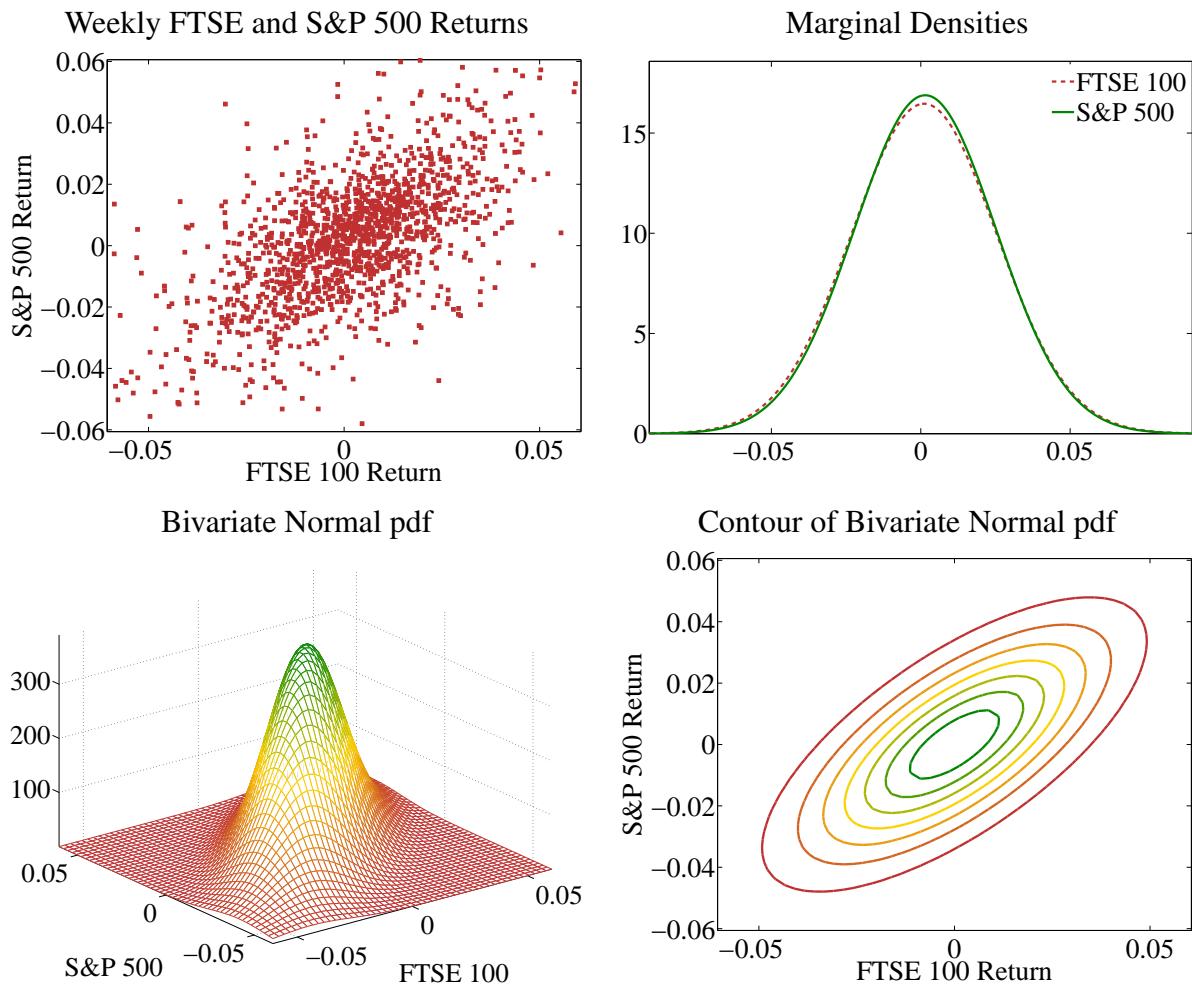


Figure 1.7: These four figures show different views of the weekly returns of the FTSE 100 and the S&P 500. The top left contains a scatter plot of the raw data. The top right shows the marginal distributions from a fit bivariate normal distribution (using maximum likelihood). The bottom two panels show two representations of the joint probability density function.

Moments

Mean	μ
Median	μ
Variance	Σ
Skewness	0
Kurtosis	3

Marginal Distribution

The marginal distribution for the first j components is

$$f_{X_1, \dots, X_j}(x_1, \dots, x_j) = (2\pi)^{-\frac{j}{2}} |\Sigma_{11}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\mathbf{x}_1 - \boldsymbol{\mu}_1)' \Sigma_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1)\right),$$

where it is assumed that the marginal distribution is that of the first j random variables¹², $\boldsymbol{\mu} = [\boldsymbol{\mu}'_1 \boldsymbol{\mu}'_2]'$ where $\boldsymbol{\mu}_1$ correspond to the first j entries, and

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma'_{12} & \Sigma_{22} \end{bmatrix}.$$

In other words, the distribution of $[X_1, \dots, X_j]'$ is $N(\boldsymbol{\mu}_1, \Sigma_{11})$. Moreover, the marginal distribution of a single element of X is $N(\mu_i, \sigma_i^2)$ where μ_i is the i^{th} element of $\boldsymbol{\mu}$ and σ_i^2 is the i^{th} diagonal element of Σ .

Conditional Distribution

The conditional probability of X_1 given $X_2 = \mathbf{x}_2$ is

$$N(\boldsymbol{\mu}_1 + \boldsymbol{\beta}'(\mathbf{x}_2 - \boldsymbol{\mu}_2), \Sigma_{11} - \boldsymbol{\beta}' \Sigma_{22} \boldsymbol{\beta})$$

where $\boldsymbol{\beta} = \Sigma_{22}^{-1} \Sigma'_{12}$.

When X is a bivariate normal random variable,

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}\right),$$

the conditional distribution is

$$X_1 | X_2 = x_2 \sim N\left(\boldsymbol{\mu}_1 + \frac{\sigma_{12}}{\sigma_2^2} (x_2 - \boldsymbol{\mu}_2), \sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2}\right),$$

where the variance can be seen to always be positive since $\sigma_1^2 \sigma_2^2 \geq \sigma_{12}^2$ by the Cauchy-Schwarz inequality (see 1.15).

¹²Any two variables can be reordered in a multivariate normal by swapping their means and reordering the corresponding rows and columns of the covariance matrix.

Notes

The multivariate Normal has a number of novel and useful properties:

- A standard multivariate normal has $\mu = \mathbf{0}$ and $\Sigma = \mathbf{I}_n$.
- If the covariance between elements i and j equals zero (so that $\sigma_{ij} = 0$), they are independent.
- For the normal, zero covariance (or correlation) implies independence. This is not true of most other multivariate random variables.
- Weighted sums of multivariate normal random variables are normal. In particular if \mathbf{c} is a n by 1 vector of weights, then $Y = \mathbf{c}'\mathbf{X}$ is normal with mean $\mathbf{c}'\mu$ and variance $\mathbf{c}'\Sigma\mathbf{c}$.

1.4 Expectations and Moments

Expectations and moments are (non-random) functions of random variables that are useful in both understanding properties of random variables – e.g. when comparing the dispersion between two distributions – and when estimating parameters using a technique known as the method of moments (see Chapter 1).

1.4.1 Expectations

The expectation is the value, on average, of a random variable (or function of a random variable). Unlike common English language usage, where one's expectation is not well defined (e.g. could be the mean or the mode, another measure of the tendency of a random variable), the expectation in a probabilistic sense *always* averages over the possible values weighting by the probability of observing each value. The form of an expectation in the discrete case is particularly simple.

Definition 1.36 (Expectation of a Discrete Random Variable). The expectation of a discrete random variable, defined $E[X] = \sum_{x \in R(X)} xf(x)$, exists if and only if $\sum_{x \in R(X)} |x| f(x) < \infty$.

When the range of X is finite then the expectation always exists. When the range is infinite, such as when a random variable takes on values in the range $0, 1, 2, \dots$, the probability mass function must be sufficiently small for large values of the random variable in order for the expectation to exist.¹³ Expectations of continuous random variables are virtually identical, only replacing the sum with an integral.

Definition 1.37 (Expectation of a Continuous Random Variable). The expectation of a continuous random variable, defined $E[X] = \int_{-\infty}^{\infty} xf(x) dx$, exists if and only if $\int_{-\infty}^{\infty} |x| f(x) dx < \infty$.

The existence of an expectation is a somewhat difficult concept. For continuous random variables, expectations may not exist if the probability of observing an arbitrarily large value (in the absolute sense) is very high. For example, in a Student's t distribution when the degree of freedom parameter v is 1 (also known as a Cauchy distribution), the probability of observing a value with size $|x|$ is

¹³An expectation is said to be nonexistent when the sum converges to $\pm\infty$ or oscillates. The use of the $|x|$ in the definition of existence is to rule out both the $-\infty$ and the oscillating cases.

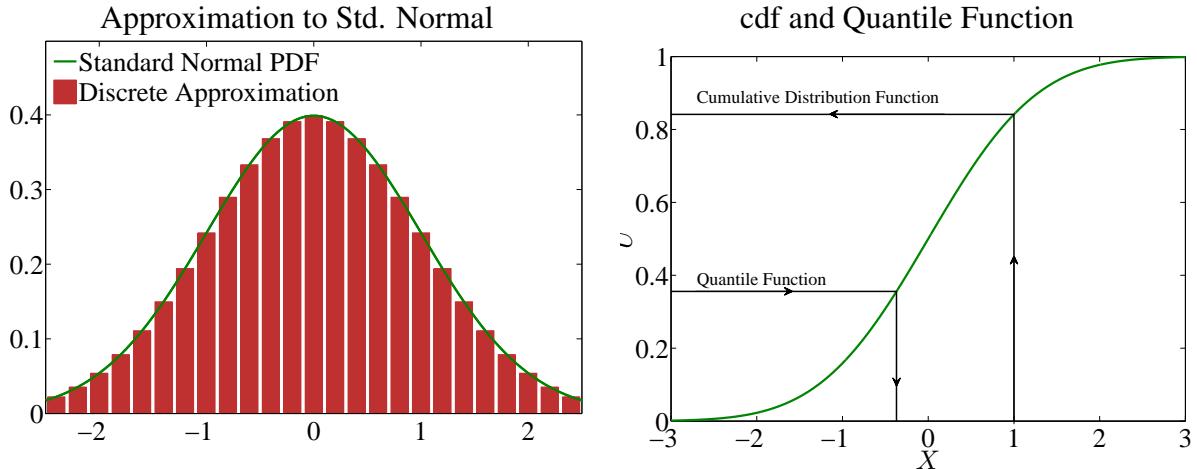


Figure 1.8: The left panel shows a standard normal and a discrete approximation. Discrete approximations are useful for approximating integrals in expectations. The right panel shows the relationship between the quantile function and the cdf.

proportional to x^{-1} for large x (in other words, $f(x) \propto cx^{-1}$) so that $xf(x) \approx c$ for large x . The range is unbounded, and so the integral of a constant, even if very small, will not converge, and so the expectation does not exist. On the other hand, when a random variable is bounded, its expectation always exists.

Theorem 1.9 (Expectation Existence for Bounded Random Variables). *If $|x| < c$ for all $x \in R(X)$, then $E[X]$ exists.*

The expectation operator, $E[\cdot]$ is generally defined for arbitrary functions of a random variable, $g(x)$. In practice, $g(x)$ takes many forms – x , x^2 , x^p for some p , $\exp(x)$ or something more complicated. Discrete and continuous expectations are closely related. Figure 1.8 shows a standard normal along with a discrete approximation where each bin has a width of 0.20 and the height is based on the pdf value at the mid-point of the bin. Treating the normal as a discrete distribution based on this approximation would provide reasonable approximations to the correct (integral) expectations.

Definition 1.38 (Expectation of a Function of Random Variable). The expectation of a random variable defined as a function of X , $Y = g(x)$, is $E[Y] = E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx$ exists if and only if $\int_{-\infty}^{\infty} |g(x)| dx < \infty$.

When $g(x)$ is either concave or convex, Jensen's inequality provides a relationship between the expected value of the function and the function of the expected value of the underlying random variable.

Theorem 1.10 (Jensen's Inequality). *If $g(\cdot)$ is a continuous convex function on an open interval containing the range of X , then $E[g(X)] \geq g(E[X])$. Similarly, if $g(\cdot)$ is a continuous concave function on an open interval containing the range of X , then $E[g(X)] \leq g(E[X])$.*

The inequalities become strict if the functions are strictly convex (or concave) as long as X is not degenerate.¹⁴ Jensen's inequality is common in economic applications. For example, standard utility

¹⁴A degenerate random variable has probability 1 on a single point, and so is not meaningfully random.

functions ($U(\cdot)$) are assumed to be concave which reflects the idea that marginal utility ($U'(\cdot)$) is decreasing in consumption (or wealth). Applying Jensen's inequality shows that if consumption is random, then $E[U(c)] < U(E[c])$ – in other words, the economic agent is worse off when facing uncertain consumption. Convex functions are also commonly encountered, for example in option pricing or in (production) cost functions. The expectations operator has a number of simple and useful properties:

- If c is a constant, then $E[c] = c$. This property follows since the expectation is an integral against a probability density which integrates to unity.
- If c is a constant, then $E[cX] = cE[X]$. This property follows directly from passing the constant out of the integral in the definition of the expectation operator.
- The expectation of the sum is the sum of the expectations,

$$E \left[\sum_{i=1}^k g_i(X) \right] = \sum_{i=1}^k E[g_i(X)].$$

This property follows directly from the distributive property of multiplication.

- If a is a constant, then $E[a + X] = a + E[X]$. This property also follows from the distributive property of multiplication.
- $E[f(X)] = f(E[X])$ when $f(x)$ is affine (i.e. $f(x) = a + bx$ where a and b are constants). For general non-linear functions, it is usually the case that $E[f(X)] \neq f(E[X])$ when X is non-degenerate.
- $E[X^p] \neq E[X]^p$ except when $p = 1$ when X is non-degenerate.

These rules are used throughout financial economics when studying random variables and functions of random variables.

The expectation of a function of a multivariate random variable is similarly defined, only integrating across all dimensions.

Definition 1.39 (Expectation of a Multivariate Random Variable). Let (X_1, X_2, \dots, X_n) be a continuously distributed n -dimensional multivariate random variable with joint density function $f(x_1, x_2, \dots, x_n)$. The expectation of $Y = g(X_1, X_2, \dots, X_n)$ is defined as

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} g(x_1, x_2, \dots, x_n) f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n. \quad (1.24)$$

It is straight forward to see that rule that the expectation of the sum is the sum of the expectation carries over to multivariate random variables, and so

$$E \left[\sum_{i=1}^n g_i(X_1, \dots, X_n) \right] = \sum_{i=1}^n E[g_i(X_1, \dots, X_n)].$$

Additionally, taking $g_i(X_1, \dots, X_n) = X_i$, we have $E[\sum_{i=1}^n X_i] = \sum_{i=1}^n E[X_i]$.

1.4.2 Moments

Moments are expectations of particular functions of a random variable, typically $g(x) = x^s$ for $s = 1, 2, \dots$, and are often used to compare distributions or to estimate parameters.

Definition 1.40 (Noncentral Moment). The r^{th} noncentral moment of a continuous random variable X is defined

$$\mu'_r \equiv E[X^r] = \int_{-\infty}^{\infty} x^r f(x) dx \quad (1.25)$$

for $r = 1, 2, \dots$

The first non-central moment is the average, or mean, of the random variable.

Definition 1.41 (Mean). The first non-central moment of a random variable X is called the mean of X and is denoted μ .

Central moments are similarly defined, only centered around the mean.

Definition 1.42 (Central Moment). The r^{th} central moment of a random variables X is defined

$$\mu_r \equiv E[(X - \mu)^r] = \int_{-\infty}^{\infty} (x - \mu)^r f(x) dx \quad (1.26)$$

for $r = 2, 3, \dots$

Aside from the first moment, references to “moments” refer to central moments. Moments may not exist if a distribution is sufficiently heavy-tailed. However, if the r^{th} moment exists, then any moment of lower order must also exist.

Theorem 1.11 (Lesser Moment Existence). If μ'_r exists for some r , then μ'_s exists for $s \leq r$. Moreover, for any r , μ'_r exists if and only if μ_r exists.

Central moments are used to describe a distribution since they are invariant to changes in the mean. The second central moment is known as the variance.

Definition 1.43 (Variance). The second central moment of a random variable X , $E[(X - \mu)^2]$ is called the variance and is denoted σ^2 or equivalently $V[X]$.

The variance operator ($V[\cdot]$) also has a number of useful properties.

- If c is a constant, then $V[c] = 0$.
- If c is a constant, then $V[cX] = c^2 V[X]$.
- If a is a constant, then $V[a + X] = V[X]$.
- The variance of the sum is the sum of the variances plus twice all of the covariances^a,

$$V\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n V[X_i] + 2 \sum_{j=1}^n \sum_{k=j+1}^n \text{Cov}[X_j, X_k]$$

^aSee Section 1.4.7 for more on covariances.

The variance is a measure of dispersion, although the square root of the variance, known as the standard deviation, is typically more useful.¹⁵

Definition 1.44 (Standard Deviation). The square root of the variance is known as the standard deviations and is denoted σ or equivalently $\text{std}(X)$.

The standard deviation is a more meaningful measure than the variance since its *units* are the same as the mean (and random variable). For example, suppose X is the return on the stock market next year, and that the mean of X is 8% and the standard deviation is 20% (the variance is .04). The mean and standard deviation are both measured as the percentage change in investment, and so can be directly compared, such as in the Sharpe ratio (Sharpe, 1994). Applying the properties of the expectation operator and variance operator, it is possible to define a studentized (or standardized) random variable.

Definition 1.45 (Studentization). Let X be a random variable with mean μ and variance σ^2 , then

$$Z = \frac{x - \mu}{\sigma} \quad (1.27)$$

is a studentized version of X (also known as standardized). Z has mean 0 and variance 1.

Standard deviation also provides a bound on the probability which can lie in the tail of a distribution, as shown in Chebyshev's inequality.

Theorem 1.12 (Chebyshev's Inequality). $\Pr[|x - \mu| \geq k\sigma] \leq 1/k^2$ for $k > 0$.

Chebyshev's inequality is useful in a number of contexts. One of the most useful is in establishing consistency in any an estimator which has a variance that tends to 0 as the sample size diverges.

The third central moment does not have a specific name, although it is called the skewness when standardized by the scaled variance.

Definition 1.46 (Skewness). The third central moment, standardized by the second central moment raised to the power 3/2,

$$\frac{\mu_3}{(\sigma^2)^{\frac{3}{2}}} = \frac{\mathbb{E}[(X - \mathbb{E}[X])^3]}{\mathbb{E}[(X - \mathbb{E}[X])^2]^{\frac{3}{2}}} = \mathbb{E}[Z^3] \quad (1.28)$$

is defined as the skewness where Z is a studentized version of X .

The skewness is a general measure of asymmetry, and is 0 for symmetric distribution (assuming the third moment exists). The normalized fourth central moment is known as the kurtosis.

Definition 1.47 (Kurtosis). The fourth central moment, standardized by the squared second central moment,

$$\frac{\mu_4}{(\sigma^2)^2} = \frac{\mathbb{E}[(X - \mathbb{E}[X])^4]}{\mathbb{E}[(X - \mathbb{E}[X])^2]^2} = \mathbb{E}[Z^4] \quad (1.29)$$

is defined as the kurtosis and is denoted κ where Z is a studentized version of X .

¹⁵The standard deviation is occasionally confused for the standard error. While both are square roots of variances, the standard deviation refers to deviation in a random variable while the standard error is reserved for parameter estimators.

Kurtosis measures of the chance of observing a large (and absolute terms) value, and is often expressed as excess kurtosis.

Definition 1.48 (Excess Kurtosis). The kurtosis of a random variable minus the kurtosis of a normal random variable, $\kappa - 3$, is known as excess kurtosis.

Random variables with a positive excess kurtosis are often referred to as heavy-tailed.

1.4.3 Related Measures

While moments are useful in describing the properties of a random variable, other measures are also commonly encountered. The median is an alternative measure of central tendency.

Definition 1.49 (Median). Any number m satisfying $\Pr(X \leq m) = 0.5$ and $\Pr(X \geq m) = 0.5$ is known as the median of X .

The median measures the point where 50% of the distribution lies on either side (it may not be unique), and is just a particular quantile. The median has a few advantages over the mean, and in particular, it is less affected by outliers (e.g. the difference between mean and median income) and it always exists (the mean doesn't exist for very heavy-tailed distributions).

The interquartile range uses quartiles¹⁶ to provide an alternative measure of dispersion than standard deviation.

Definition 1.50 (Interquartile Range). The value $q_{.75} - q_{.25}$ is known as the interquartile range.

The mode complements the mean and median as a measure of central tendency. A mode is a local maximum of a density.

Definition 1.51 (Mode). Let X be a random variable with density function $f(x)$. A point c where $f(x)$ attains a maximum is known as a mode.

Distributions can be unimodal or multimodal.

Definition 1.52 (Unimodal Distribution). Any random variable which has a single, unique mode is called unimodal.

Note that modes in a multimodal distribution do not necessarily have to have equal probability.

Definition 1.53 (Multimodal Distribution). Any random variable which has more than one mode is called multimodal.

Figure 1.9 shows a number of distributions. The distributions depicted in the top panels are all unimodal. The distributions in the bottom pane are mixtures of normals, meaning that with probability p random variables come from one normal, and with probability $1 - p$ they are drawn from the other. Both mixtures of normals are multimodal.

¹⁶Other tiles include terciles (3), quartiles (4), quintiles (5), deciles (10) and percentiles (100). In all cases the bin ends are $[(i - 1/m), i/m]$ where m is the number of bins and $i = 1, 2, \dots, m$.

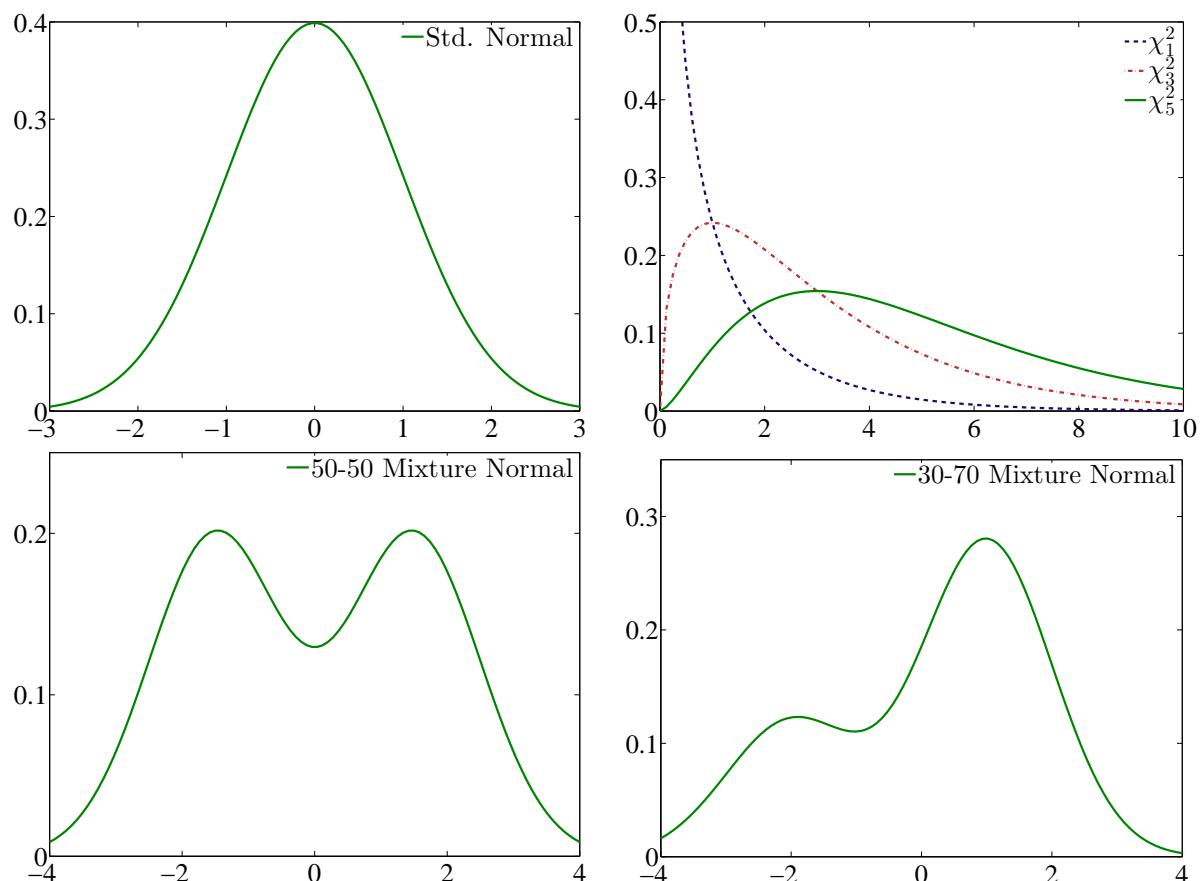


Figure 1.9: These four figures show two unimodal (upper panels) and two multimodal (lower panels) distributions. The upper left is a standard normal density. The upper right shows three χ^2 densities for $v = 1, 3$ and 5 . The lower panels contain mixture distributions of 2 normals – the left is a 50-50 mixture of $N(-1, 1)$ and $N(1, 1)$ and the right is a 30-70 mixture of $N(-2, 1)$ and $N(1, 1)$.

1.4.4 Multivariate Moments

Other moment definitions are only meaningful when studying 2 or more random variables (or an n -dimensional random variable). When applied to a vector or matrix, the expectations operator applies element-by-element. For example, if X is an n -dimensional random variable,

$$\mathbb{E}[X] = \mathbb{E} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_n] \end{bmatrix}. \quad (1.30)$$

Covariance is a measure which captures the tendency of two variables to move together in a linear sense.

Definition 1.54 (Covariance). The covariance between two random variables X and Y is defined

$$\text{Cov}[X, Y] = \sigma_{XY} = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]. \quad (1.31)$$

Covariance can be alternatively defined using the joint product moment and the product of the means.

Theorem 1.13 (Alternative Covariance). *The covariance between two random variables X and Y can be equivalently defined*

$$\sigma_{XY} = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]. \quad (1.32)$$

Inverting the covariance expression shows that no covariance is sufficient to ensure that the expectation of a product is the product of the expectations.

Theorem 1.14 (Zero Covariance and Expectation of Product). *If X and Y have $\sigma_{XY} = 0$, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.*

The previous result follows directly from the definition of covariance since $\sigma_{XY} = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$. In financial economics, this result is often applied to products of random variables so that the mean of the product can be directly determined by knowledge of the mean of each variable and the covariance between the two. For example, when studying consumption based asset pricing, it is common to encounter terms involving the expected value of consumption growth times the pricing kernel (or stochastic discount factor) – in many cases the full joint distribution of the two is intractable although the mean and covariance of the two random variables can be determined.

The Cauchy-Schwarz inequality is a version of the triangle inequality and states that the expectation of the squared product is less than the product of the squares.

Theorem 1.15 (Cauchy-Schwarz Inequality). $\mathbb{E}[(XY)^2] \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$.

Example 1.32. When X is an n -dimensional random variable, it is useful to assemble the variances and covariances into a covariance matrix.

Definition 1.55 (Covariance Matrix). The covariance matrix of an n -dimensional random variable X is defined

$$\text{Cov}[X] = \Sigma = E[(X - E[X])(X - E[X])'] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \vdots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \sigma_n^2 \end{bmatrix}$$

where the i^{th} diagonal element contains the variance of X_i (σ_i^2) and the element in position (i, j) contains the covariance between X_i and X_j (σ_{ij}).

When X is composed of two sub-vectors, a block form of the covariance matrix is often convenient.

Definition 1.56 (Block Covariance Matrix). Suppose X_1 is an n_1 -dimensional random variable and X_2 is an n_2 -dimensional random variable. The block covariance matrix of $X = [X'_1 X'_2]'$ is

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma'_{12} & \Sigma_{22} \end{bmatrix} \quad (1.33)$$

where Σ_{11} is the n_1 by n_1 covariance of X_1 , Σ_{22} is the n_2 by n_2 covariance of X_2 and Σ_{12} is the n_1 by n_2 covariance matrix between X_1 and X_2 and element (i, j) equal to $\text{Cov}[X_{1,i}, X_{2,j}]$.

A standardized version of covariance is often used to produce a scale-free measure.

Definition 1.57 (Correlation). The correlation between two random variables X and Y is defined

$$\text{Corr}[X, Y] = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}. \quad (1.34)$$

Additionally, the correlation is always in the interval $[-1, 1]$, which follows from the Cauchy-Schwarz inequality.

Theorem 1.16. If X and Y are independent random variables, then $\rho_{XY} = 0$ as long as σ_X^2 and σ_Y^2 exist.

It is important to note that the converse of this statement is not true – that is, a lack of correlation does not imply that two variables are independent. In general, a correlation of 0 only implies independence when the variables are multivariate normal.

Example 1.33. Suppose X and Y have $\rho_{XY} = 0$, then X and Y are not necessarily independent. Suppose X is a discrete uniform random variable taking values in $\{-1, 0, 1\}$ and $Y = X^2$, so that $\sigma_X^2 = 2/3$, $\sigma_Y^2 = 2/9$ and $\sigma_{XY} = 0$. While X and Y are uncorrelated, they are clearly not independent, since when the random variable Y takes the value 1, X must be 0.

The corresponding correlation matrix can be assembled. Note that a correlation matrix has 1s on the diagonal and values bounded by $[-1, 1]$ on the off-diagonal positions.

Definition 1.58 (Correlation Matrix). The correlation matrix of an n -dimensional random variable X is defined

$$(\Sigma \odot \mathbf{I}_n)^{-\frac{1}{2}} \Sigma (\Sigma \odot \mathbf{I}_n)^{-\frac{1}{2}} \quad (1.35)$$

where the i, j^{th} element has the form $\sigma_{X_i X_j} / (\sigma_{X_i} \sigma_{X_j})$ when $i \neq j$ and 1 when $i = j$.

1.4.5 Conditional Expectations

Conditional expectations are similar to other forms of expectations only using conditional densities in place of joint or marginal densities. Conditional expectations essentially treat one of the variables (in a bivariate random variable) as constant.

Definition 1.59 (Bivariate Conditional Expectation). Let X be a continuous bivariate random variable comprised of X_1 and X_2 . The conditional expectation of X_1 given X_2

$$E[g(X_1)|X_2 = x_2] = \int_{-\infty}^{\infty} g(x_1) f(x_1|x_2) dx_1 \quad (1.36)$$

where $f(x_1|x_2)$ is the conditional probability density function of X_1 given X_2 .¹⁷

In many cases, it is useful to avoid specifying a specific value for X_2 in which case $E[X_1|X_1]$ will be used. Note that $E[X_1|X_2]$ will typically be a function of the random variable X_2 .

Example 1.34. Suppose X is a bivariate normal distribution with components X_1 and X_2 , $\mu = [\mu_1 \mu_2]'$ and

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix},$$

then $E[X_1|X_2 = x_2] = \mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(x_2 - \mu_2)$. This follows from the conditional density of a bivariate random variable.

The law of iterated expectations uses conditional expectations to show that the conditioning does not affect the final result of taking expectations – in other words, the order of taking expectations does not matter.

Theorem 1.17 (Bivariate Law of Iterated Expectations). *Let X be a continuous bivariate random variable comprised of X_1 and X_2 . Then $E[E[g(X_1)|X_2]] = E[g(X_1)]$.*

The law of iterated expectations follows from basic properties of an integral since the order of integration does not matter as long as all integrals are taken.

Example 1.35. Suppose X is a bivariate normal distribution with components X_1 and X_2 , $\mu = [\mu_1 \mu_2]'$ and

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix},$$

then $E[X_1] = \mu_1$ and

$$\begin{aligned} E[E[X_1|X_2]] &= E\left[\mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(X_2 - \mu_2)\right] \\ &= \mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(E[X_2] - \mu_2) \\ &= \mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(\mu_2 - \mu_2) \\ &= \mu_1. \end{aligned}$$

¹⁷A conditional expectation can also be defined in a natural way for functions of X_1 given $X_2 \in B$ where $\Pr(X_2 \in B) > 0$.

When using conditional expectations, any random variable conditioned on behaves “as-if” non-random (in the conditional expectation), and so $E[E[X_1X_2|X_2]] = E[X_2E[X_1|X_2]]$. This is a very useful tool when combined with the law of iterated expectations when $E[X_1|X_2]$ is a known function of X_2 .

Example 1.36. Suppose X is a bivariate normal distribution with components X_1 and X_2 , $\mu = \mathbf{0}$ and

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix},$$

then

$$\begin{aligned} E[X_1X_2] &= E[E[X_1X_2|X_2]] \\ &= E[X_2E[X_1|X_2]] \\ &= E\left[X_2\left(\frac{\sigma_{12}}{\sigma_2^2}X_2\right)\right] \\ &= \frac{\sigma_{12}}{\sigma_2^2}E[X_2^2] \\ &= \frac{\sigma_{12}}{\sigma_2^2}(\sigma_2^2) \\ &= \sigma_{12}. \end{aligned}$$

One particularly useful application of conditional expectations occurs when the conditional expectation is known and constant, so that $E[X_1|X_2] = c$.

Example 1.37. Suppose X is a bivariate random variable composed of X_1 and X_2 and that $E[X_1|X_2] = c$. Then $E[X_1] = c$ since

$$\begin{aligned} E[X_1] &= E[E[X_1|X_2]] \\ &= E[c] \\ &= c. \end{aligned}$$

Conditional expectations can be taken for general n -dimensional random variables, and the law of iterated expectations holds as well.

Definition 1.60 (Conditional Expectation). Let X be a n -dimensional random variable and partition the first j ($1 \leq j < n$) elements of X into X_1 , and the remainder into X_2 so that $X = [X'_1 X'_2]'$. The conditional expectation of $g(X_1)$ given $X_2 = \mathbf{x}_2$

$$E[g(X_1)|X_2 = \mathbf{x}_2] = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} g(x_1, \dots, x_j) f(x_1, \dots, x_j | \mathbf{x}_2) dx_j \dots dx_1 \quad (1.37)$$

where $f(x_1, \dots, x_j | \mathbf{x}_2)$ is the conditional probability density function of X_1 given $X_2 = \mathbf{x}_2$.

The law of iterated expectations also holds for arbitrary partitions as well.

Theorem 1.18 (Law of Iterated Expectations). *Let X be a n -dimensional random variable and partition the first j ($1 \leq j < n$) elements of X into X_1 , and the remainder into X_2 so that $X = [X'_1 X'_2]'$. Then $E[E[g(X_1)|X_2]] = E[g(X_1)]$. The law of iterated expectations is also known as the law of total expectations.*

Full multivariate conditional expectations are extremely common in time series. For example, when using daily data, there are over 30,000 observations of the Dow Jones Industrial Average available to model. Attempting to model the full joint distribution would be a formidable task. On the other hand, modeling the conditional expectation (or conditional mean) of the final observation, conditioning on those observations in the past, is far simpler.

Example 1.38. Suppose $\{X_t\}$ is a sequence of random variables where X_t comes after X_{t-j} for $j \geq 1$. The conditional conditional expectation of X_t given its past is

$$\mathbb{E}[X_t | X_{t-1}, X_{t-2}, \dots].$$

Example 1.39. Let $\{\varepsilon_t\}$ be a sequence of independent, identically distributed random variables with mean 0 and variance $\sigma^2 < \infty$. Define $X_0 = 0$ and $X_t = X_{t-1} + \varepsilon_t$. X_t is a random walk, and $\mathbb{E}[X_t | X_{t-1}] = X_{t-1}$.

This leads naturally to the definition of a martingale, which is an important concept in financial economics which related to efficient markets.

Definition 1.61 (Martingale). If $\mathbb{E}[X_{t+j} | X_{t-1}, X_{t-2}, \dots] = X_{t-1}$ for all $j \geq 0$ and $\mathbb{E}[|X_t|] < \infty$, both holding for all t , then $\{X_t\}$ is a martingale. Similarly, if $\mathbb{E}[X_{t+j} - X_{t-1} | X_{t-1}, X_{t-2}, \dots] = 0$ for all $j \geq 0$ and $\mathbb{E}[|X_t|] < \infty$, both holding for all t , then $\{X_t\}$ is a martingale.

1.4.6 Conditional Moments

All moments can be transformed made conditional by integrating against the conditional probability density function. For example, the (unconditional) mean becomes the conditional mean, and the variance becomes a conditional variance.

Definition 1.62 (Conditional Variance). The variance of a random variable X conditional on another random variable Y is

$$\begin{aligned} \text{V}[X|Y] &= \mathbb{E}\left[(X - \mathbb{E}[X|Y])^2 | Y\right] \\ &= \mathbb{E}[X^2|Y] - \mathbb{E}[X|Y]^2. \end{aligned} \tag{1.38}$$

The two definitions of conditional variance are identical to those of the (unconditional) variance where the (unconditional) expectation has been replaced by a conditional expectation. Conditioning can be used to compute higher-order moments as well.

Definition 1.63 (Conditional Moment). The r^{th} central moment of a random variables X conditional on another random variable Y is defined

$$\mu_r \equiv \mathbb{E}\left[(X - \mathbb{E}[X|Y])^r | Y\right] \tag{1.39}$$

for $r = 2, 3, \dots$

Combining the conditional expectation and the conditional variance leads to the law of total variance.

Theorem 1.19. *The variance of a random variable X can be decomposed into the variance of the conditional expectation plus the expectation of the conditional variance,*

$$\text{V}[X] = \text{V}[\text{E}[X|Y]] + \text{E}[\text{V}[X|Y]]. \quad (1.40)$$

The law of total variance shows that the total variance of a variable can be decomposed into the variability of the conditional mean plus the average of the conditional variance. This is a useful decomposition for time-series.

Independence can also be defined conditionally.

Definition 1.64 (Conditional Independence). Two random variables X_1 and X_2 are conditionally independent, conditional on Y , if

$$f(x_1, x_2|y) = f_1(x_1|y)f_2(x_2|y).$$

Note that random variables that are conditionally independent are not necessarily unconditionally independent.

Example 1.40. Suppose X is a trivariate normal random variable with mean $\mathbf{0}$ and covariance

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix}$$

and define $Y_1 = x_1 + x_3$ and $Y_2 = x_2 + x_3$. Then Y_1 and Y_2 are correlated bivariate normal with mean $\mathbf{0}$ and covariance

$$\Sigma_Y = \begin{bmatrix} \sigma_1^2 + \sigma_3^2 & \sigma_3^2 \\ \sigma_3^2 & \sigma_2^2 + \sigma_3^2 \end{bmatrix},$$

but the joint distribution of Y_1 and Y_2 given X_3 is bivariate normal with mean $\mathbf{0}$ and

$$\Sigma_{Y|X_3} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

and so Y_1 and Y_2 are independent conditional on X_3 .

Other properties of unconditionally independent random variables continue to hold for conditionally independent random variables. For example, when X_1 and X_2 are independent conditional on X_3 , then the conditional covariance between X_1 and X_2 is 0 (as is the conditional correlation), and $\text{E}[\text{E}[X_1 X_2|X_3]] = \text{E}[\text{E}[X_1|X_3]\text{E}[X_2|X_3]]$ – that is, the conditional expectation of the product is the product of the conditional expectations.

1.4.7 Vector and Matrix Forms

Vector and matrix forms are particularly useful in finance since portfolios are often of interest where the underlying random variables are the individual assets and the combination vector is the vector of portfolio weights.

Theorem 1.20. *Let $Y = \sum_{i=1}^n c_i X_i$ where $c_i, i = 1, \dots, n$ are constants. Then $\text{E}[Y] = \sum_{i=1}^n c_i \text{E}[X_i]$. In matrix notation, $Y = \mathbf{c}' \mathbf{x}$ where \mathbf{c} is an n by 1 vector and $\text{E}[Y] = \mathbf{c}' \text{E}[\mathbf{x}]$.*

The variance of the sum is the weighted sum of the variance plus all of the covariances.

Theorem 1.21. Let $Y = \sum_{i=1}^n c_i X_i$ where c_i are constants. Then

$$V[Y] = \sum_{i=1}^n c_i^2 V[X_i] + 2 \sum_{j=1}^n \sum_{k=j+1}^n c_j c_k \text{Cov}[X_i, X_j] \quad (1.41)$$

or equivalently

$$\sigma_Y^2 = \sum_{i=1}^n c_i^2 \sigma_{X_i}^2 + 2 \sum_{j=1}^n \sum_{k=j+1}^n c_j c_k \sigma_{X_j X_k}.$$

This result can be equivalently expressed in vector-matrix notation.

Theorem 1.22. Let \mathbf{c} in an n by 1 vector and let X by an n -dimensional random variable with covariance Σ . Define $Y = \mathbf{c}' \mathbf{x}$. The variance of Y is $\sigma_Y^2 = \mathbf{c}' \text{Cov}[X] \mathbf{c} = \mathbf{c}' \Sigma \mathbf{c}$.

Note that the result holds when \mathbf{c} is replaced by a matrix \mathbf{C} .

Theorem 1.23. Let \mathbf{C} be an n by m matrix and let X be an n -dimensional random variable with mean μ_X and covariance Σ_X . Define $Y = \mathbf{C}' \mathbf{x}$. The expected value of Y is $E[Y] = \mu_Y = \mathbf{C}' E[X] = \mathbf{C}' \mu_X$ and the covariance of Y is $\Sigma_Y = \mathbf{C}' \text{Cov}[X] \mathbf{C} = \mathbf{C}' \Sigma_X \mathbf{C}$.

Definition 1.65 (Multivariate Studentization). Let X be an n -dimensional random variable with mean μ and covariance Σ , then

$$Z = \Sigma^{-\frac{1}{2}} (\mathbf{x} - \mu) \quad (1.42)$$

is a studentized version of X where $\Sigma^{\frac{1}{2}}$ is a matrix square root such as the Cholesky factor or one based on the spectral decomposition of Σ . Z has mean $\mathbf{0}$ and covariance equal to the identity matrix \mathbf{I}_n .

The final result for vectors relates quadratic forms of normals (inner-products) to χ^2 distributed random variables.

Theorem 1.24 (Quadratic Forms of Normals). Let X be an n -dimensional normal random variable with mean $\mathbf{0}$ and identity covariance \mathbf{I}_n . Then $\mathbf{x}' \mathbf{x} = \sum_{i=1}^n x_i^2 \sim \chi_n^2$.

Combining this result with studentization, when X is a general n -dimensional normal random variable with mean μ and covariance Σ ,

$$(\mathbf{x} - \mu)' \left(\Sigma^{-\frac{1}{2}} \right)' \Sigma^{-\frac{1}{2}} (\mathbf{x} - \mu)' = (\mathbf{x} - \mu)' \Sigma^{-1} (\mathbf{x} - \mu)' \sim \chi_n^2.$$

1.4.8 Monte Carlo and Numerical Integration

Expectations of functions of continuous random variables are integrals against the underlying pdf. In some cases, these integrals are analytically tractable, although in many situations integrals cannot be analytically computed and so numerical techniques are needed to compute expected values and moments.

Monte Carlo is one method to approximate an integral. Monte Carlo utilizes simulated draws from the underlying distribution and averaging to approximate integrals.

Definition 1.66 (Monte Carlo Integration). Suppose $X \sim F(\theta)$ and that it is possible to simulate a series $\{x_i\}$ from $F(\theta)$. The Monte Carlo expectation of a function $g(x)$ is defined

$$\widehat{E[g(X)]} = m^{-1} \sum_{i=1}^m g(x_i),$$

Moreover, as long as $E[|g(x)|] < \infty$, $\lim_{m \rightarrow \infty} m^{-1} \sum_{i=1}^m g(x_i) = E[g(x)]$.

The intuition behind this result follows from the properties of $\{x_i\}$. Since these are i.i.d. draws from $F(\theta)$, they will, on average, tend to appear in any interval $B \in R(X)$ in proportion to the probability $\Pr(X \in B)$. In essence, the simulated values coarsely approximating the discrete approximation shown in 1.8.

While Monte Carlo integration is a general technique, there are some important limitations. First, if the function $g(x)$ takes large values in regions where $\Pr(X \in B)$ is small, it may require a very large number of draws to accurately approximate $E[g(x)]$ since, by construction, there are unlikely to be many points in B . In practice the behavior of $h(x) = g(x)f(x)$ plays an important role in determining the appropriate sample size.¹⁸ Second, while Monte Carlo integration is technically valid for random variables with any number of dimensions, in practice it is usually only reliable when the dimension is small (typically 3 or fewer), especially when the range is unbounded ($R(X) \in \mathbb{R}^n$). When the dimension of X is large, many simulated draws are needed to visit the corners of the (joint) pdf, and if 1,000 draws are sufficient for a unidimensional problem, 1000^n may be needed to achieve the same accuracy when X has n dimensions.

Alternatively the function to be integrated can be approximated using a polygon with an easy-to-compute area, such as the rectangles approximating the normal pdf shown in figure 1.8. The quality of the approximation will depend on the resolution of the grid used. Suppose u and l are the upper and lower bounds of the integral, respectively, and that the region can be split into m intervals $l = b_0 < b_1 < \dots < b_{m-1} < b_m = u$. Then the integral of a function $h(\cdot)$ is

$$\int_l^u h(x) dx = \sum_{i=1}^m \int_{b_{i-1}}^{b_i} h(x) dx.$$

In practice, l and u may be infinite, in which case some cut-off point is required. In general, the cut-off should be chosen so that the vast majority of the probability lies between l and u ($\int_l^u f(x) dx \approx 1$).

This decomposition is combined with an area for approximating the area under h between b_{i-1} and b_i . The simplest is the rectangle method, which uses a rectangle with a height equal to the value of the function at the mid-point.

Definition 1.67 (Rectangle Method). The rectangle rule approximates the area under the curve with a rectangle and is given by

$$\int_l^u h(x) dx \approx h\left(\frac{u+l}{2}\right)(u-l).$$

The rectangle rule would be exact if the function was piece-wise flat. The trapezoid rule improves the approximation by replacing the function at the midpoint with the average value of the function and would be exact for any piece-wise linear function (including piece-wise flat functions).

¹⁸Monte Carlo integrals can also be seen as estimators, and in many cases standard inference can be used to determine the accuracy of the integral. See Chapter 1 for more details on inference and constructing confidence intervals.

Definition 1.68 (Trapezoid Method). The trapezoid rule approximates the area under the curve with a trapezoid and is given by

$$\int_l^u h(x) dx \approx \frac{h(u) + h(l)}{2} (u - l).$$

The final method is known as Simpson's rule which is based on using a quadratic approximation to the underlying function. It is exact when the underlying function is piece-wise linear or quadratic.

Definition 1.69 (Simpson's Rule). Simpson's Rule uses an approximation that would be exact if they underlying function were quadratic, and is given by

$$\int_l^u h(x) dx \approx \frac{u-l}{6} \left(h(u) + 4h\left(\frac{u+l}{2}\right) + h(l) \right).$$

Example 1.41. Consider the problem of computing the expected payoff of an option. The payoff of a call option is given by

$$c = \max(s_1 - k, 0)$$

where k is the strike price, s_1 is the stock price at expiration and s_0 is the current stock price. Suppose returns are normally distributed with mean $\mu = .08$ and standard deviation $\sigma = .20$. In this problem, $g(r) = (s_0 \exp(r) - k) I_{[s_0 \exp(r) > k]}$ where $I_{[\cdot]}$ and a binary indicator function which takes the value 1 when the argument is true, and

$$f(r) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(r-\mu)^2}{2\sigma^2}\right).$$

Combined, the function the be integrated is

$$\begin{aligned} \int_{-\infty}^{\infty} h(r) dr &= \int_{-\infty}^{\infty} g(r) f(r) dr \\ &= \int_{-\infty}^{\infty} (s_0 \exp(r) - k) I_{[s_0 \exp(r) > k]} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(r-\mu)^2}{2\sigma^2}\right) dr \end{aligned}$$

$s_0 = k = 50$ was used in all results.

All four methods were applied to the problem. The number of bins and the range of integration was varied for the analytical approximations. The number of bins ranged across $\{10, 20, 50, 1000\}$ and the integration range spanned $\{\pm 3\sigma, \pm 4\sigma, \pm 6\sigma, \pm 10\sigma\}$ and the bins were uniformly spaced along the integration range. Monte Carlo integration was also applied with $m \in \{100, 1000\}$.

All thing equal, increasing the number of bins increases the accuracy of the approximation. In this example, 50 appears to be sufficient. However, having a range which is too small produces values which differ from the correct value of 7.33. The sophistication of the method also improves the accuracy, especially when the number of nodes is small. The Monte Carlo results are also close, on average. However, the standard deviation is large, about 5%, even when 1000 draws are used, so that large errors would be commonly encountered and so many more points are needed to ensure that the integral is always accurate.

Shorter Problems

Problem 1.1. Suppose

$$\begin{bmatrix} X \\ U \end{bmatrix} \sim N \left(\begin{bmatrix} \mu_X \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & 0 \\ 0 & \sigma_U^2 \end{bmatrix} \right)$$

and $Y = 2X + U$. What is $E[Y]$ and $V[Y]$?

Problem 1.2. Show $\text{Cov}[aX + bY, cX + dY] = acV[X] + bdV[Y] + (ad + bc)\text{Cov}[X, Y]$.

Problem 1.3. Show that the two forms of the covariance,

$$E[XY] - E[X]E[Y] \text{ and } E[(X - E[X])(Y - E[Y])]$$

are equivalent when X and Y are continuous random variables.

Problem 1.4. Suppose $\{X_i\}$ is a sequence of random variables where $V[X_i] = \sigma^2$ for all i , $\text{Cov}[X_i, X_{i-1}] = \theta$ and $\text{Cov}[X_i, X_{i-j}] = 0$ for $j > 1$. What is $V[\bar{X}]$ where $\bar{X} = n^{-1} \sum_{i=1}^n X_i$?

Problem 1.5. Suppose $Y = \beta X + \varepsilon$ where $X \sim N(\mu_X, \sigma_X^2)$, $\varepsilon \sim N(0, \sigma^2)$ and X and ε are independent. What is $\text{Corr}[X, Y]$?

Problem 1.6. Prove that $E[a + bX] = a + bE[X]$ when X is a continuous random variable.

Problem 1.7. Prove that $V[a + bX] = b^2V[X]$ when X is a continuous random variable.

Problem 1.8. Prove that $\text{Cov}[a + bX, c + dY] = bd\text{Cov}[X, Y]$ when X and Y are a continuous random variables.

Problem 1.9. Prove that $V[a + bX + cY] = b^2V[X] + c^2V[Y] + 2bc\text{Cov}[X, Y]$ when X and Y are a continuous random variables.

Problem 1.10. Suppose $\{X_i\}$ is an i.i.d. sequence of random variables. Show that

$$V[\bar{X}] = V \left[\frac{1}{n} \sum_{i=1}^n X_i \right] = n^{-1} \sigma^2$$

where σ^2 is $V[X_1]$.

Problem 1.11. Prove that $\text{Corr}[a + bX, c + dY] = \text{Corr}[X, Y]$.

Problem 1.12. Suppose $\{X_i\}$ is a sequence of random variables where, for all i , $V[X_i] = \sigma^2$, $\text{Cov}[X_i, X_{i-1}] = \theta$ and $\text{Cov}[X_i, X_{i-j}] = 0$ for $j > 1$. What is $V[\bar{X}]$?

Problem 1.13. Prove that $E[a + bX|Y] = a + bE[X|Y]$ when X and Y are continuous random variables.

Problem 1.14. Suppose that $E[X|Y] = Y^2$ where Y is normally distributed with mean μ and variance σ^2 . What is $E[a + bX]$?

Problem 1.15. Suppose $E[X|Y = y] = a + by$ and $V[X|Y = y] = c + dy^2$ where Y is normally distributed with mean μ and variance σ^2 . What is $V[X]$?

Problem 1.16. Show that the law of total variance holds for a $V[X_1]$ when X is a bivariate normal with mean $\mu = [\mu_1 \mu_2]'$ and covariance

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}.$$

Longer Exercises

Exercise 1.1. Sixty percent (60%) of all traders hired by a large financial firm are rated as performing satisfactorily or better in their first-year review. Of these, 90% earned a first in financial econometrics. Of the traders who were rated as unsatisfactory, only 20% earned a first in financial econometrics.

1. What is the probability that a trader is rated as satisfactory or better given they received a first in financial econometrics?
2. What is the probability that a trader is rated as unsatisfactory given they received a first in financial econometrics?
3. Is financial econometrics a useful indicator of trader performance? Why or why not?

Exercise 1.2. Large financial firms use automated screening to detect rogue trades – those that exceed risk limits. One of your colleagues has introduced a new statistical test using the trading data that, given that a trader has exceeded her risk limit, detects this with probability 98%. It also only indicates false positives – that is non-rogue trades that are flagged as rogue – 1% of the time.

1. Assuming 99% of trades are legitimate, what is the probability that a detected trade is rogue? Explain the intuition behind this result.
2. Is this a useful test? Why or why not?
3. How low would the false positive rate have to be to have a 98% chance that a detected trade was actually rogue?

Exercise 1.3. Your corporate finance professor uses a few jokes to add levity to his lectures. Each week he tells 3 different jokes. However, he is also very busy, and so forgets week to week which jokes were used.

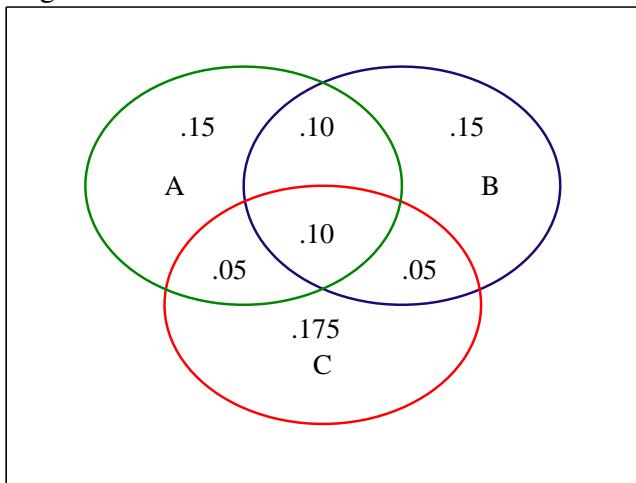
1. Assuming he has 12 jokes, what is the probability of 1 repeat across 2 consecutive weeks?
2. What is the probability of hearing 2 of the same jokes in consecutive weeks?
3. What is the probability that all 3 jokes are the same?
4. Assuming the term is 8 weeks long, and your professor has 96 jokes, what is the probability that there is no repetition during the term? Note that he remembers the jokes he gives in a particular lecture, only forgets across lectures.
5. How many jokes would your professor need to know to have a 99% chance of not repeating any in the term?

Exercise 1.4. A hedge fund company manages three distinct funds. In any given month, the probability that the return is positive is shown in the following table:

$$\begin{aligned} \Pr(r_{1,t} > 0) &= .55 & \Pr(r_{1,t} > 0 \cup r_{2,t} > 0) &= .82 \\ \Pr(r_{2,t} > 0) &= .60 & \Pr(r_{1,t} > 0 \cup r_{3,t} > 0) &= .7525 \\ \Pr(r_{3,t} > 0) &= .45 & \Pr(r_{2,t} > 0 \cup r_{3,t} > 0) &= .78 \\ \Pr(r_{2,t} > 0 \cap r_{3,t} > 0 | r_{1,t} > 0) &= .20 \end{aligned}$$

1. Are the events of “positive returns” pairwise independent?
2. Are the events of “positive returns” independent?
3. What is the probability that funds 1 and 2 have positive returns, given that fund 3 has a positive return?
4. What is the probability that at least one fund will have a positive return in any given month?

Exercise 1.5. Suppose the probabilities of three events, A , B and C are as depicted in the following diagram:



1. Are the three events pairwise independent?
2. Are the three events independent?
3. What is $\Pr(A \cap B)$?
4. What is $\Pr(A \cap B | C)$?
5. What is $\Pr(C | A \cap B)$?
6. What is $\Pr(C | A \cup B)$?

Exercise 1.6. At a small high-frequency hedge fund, two competing algorithms produce trades. Algorithm α produces 80 trades per second and 5% lose money. Algorithm β produces 20 trades per second but only 1% lose money. Given the last trade lost money, what is the probability it was produced by algorithm β ?

Exercise 1.7. Suppose $f(x, y) = 2 - x - y$ where $x \in [0, 1]$ and $y \in [0, 1]$.

1. What is $\Pr(X > .75 \cap Y > .75)$?
2. What is $\Pr(X + Y > 1.5)$?
3. Show formally whether X and Y are independent.

4. What is $\Pr(Y < .5|X = x)$?

Exercise 1.8. Suppose $f(x, y) = xy$ for $x \in [0, 1]$ and $y \in [0, 2]$.

1. What is the joint cdf?
2. What is $\Pr(X < 0.5 \cap Y < 1)$?
3. What is the marginal cdf of X ? What is $\Pr(X < 0.5)$?
4. What is the marginal density of X ?
5. Are X and Y independent?

Exercise 1.9. Suppose $F(x) = 1 - p^{x+1}$ for $x \in [0, 1, 2, \dots]$ and $p \in (0, 1)$.

1. Find the pmf.
2. Verify that the pmf is valid.
3. What is $\Pr(X \leq 8)$ if $p = .75$?
4. What is $\Pr(X \leq 1)$ given $X \leq 8$?

Exercise 1.10. A firm producing widgets has a production function $q(L) = L^{0.5}$ where L is the amount of labor. Sales prices fluctuate randomly and can be \$10 (20%), \$20 (50%) or \$30 (30%). Labor prices also vary and can be \$1 (40%), 2 (30%) or 3 (30%). The firm always maximizes profits after seeing both sales prices and labor prices.

1. Define the distribution of profits possible?
2. What is the probability that the firm makes at least \$100?
3. Given the firm makes a profit of \$100, what is the probability that the profit is over \$200?

Exercise 1.11. A fund manager tells you that her fund has non-linear returns as a function of the market and that his return is $r_{i,t} = 0.02 + 2r_{m,t} - 0.5r_{m,t}^2$ where $r_{i,t}$ is the return on the fund and $r_{m,t}$ is the return on the market.

1. She tells you her expectation of the market return this year is 10%, and that her fund will have an expected return of 22%. Can this be?
2. At what variance is would the expected return on the fund be negative?

Exercise 1.12. For the following densities, find the mean (if it exists), variance (if it exists), median and mode, and indicate whether the density is symmetric.

1. $f(x) = 3x^2$ for $x \in [0, 1]$
2. $f(x) = 2x^{-3}$ for $x \in [1, \infty)$

$$3. f(x) = [\pi(1+x^2)]^{-1} \text{ for } x \in (-\infty, \infty)$$

$$4. f(x) = \binom{4}{x} \cdot 2^x \cdot 8^{4-x} \text{ for } x \in \{0, 1, 2, 3, 4\}$$

Exercise 1.13. The daily price of a stock has an average value of £2. Then then $\Pr(X > 10) < .2$ where X denotes the price of the stock. True or false?

Exercise 1.14. An investor can invest in stocks or bonds which have expected returns and covariances as

$$\mu = \begin{bmatrix} .10 \\ .03 \end{bmatrix}, \Sigma = \begin{bmatrix} .04 & -.003 \\ -.003 & .0009 \end{bmatrix}$$

where stocks are the first component.

1. Suppose the investor has £1,000 to invest and splits the investment evenly. What is the expected return, standard deviation, variance and Sharpe Ratio (μ/σ) for the investment?
2. Now suppose the investor seeks to maximize her expected utility where her utility is defined is defined in terms of her portfolio return, $U(r) = E[r] - .01V[r]$. How much should she invest in each asset?

Exercise 1.15. Suppose $f(x) = (1-p)^x p$ for $x \in (0, 1, \dots)$ and $p \in (0, 1]$. Show that a random variable from the distribution is “memoryless” in the sense that $\Pr(X \geq s+r | X \geq r) = \Pr(X \geq s)$. In other words, the probability of surviving s or more periods is the same whether starting at 0 or after having survived r periods.

Exercise 1.16. Your Economics professor offers to play a game with you. You pay £1,000 to play and your Economics professor will flip a fair coin and pay you 2^x where x is the number of tries required for the coin to show heads.

1. What is the pmf of X ?
2. What is the expected payout from this game?

Exercise 1.17. Consider the roll of a fair pair of dice where a roll of a 7 or 11 pays $2x$ and anything else pays $-x$ where x is the amount bet. Is this game fair?

Exercise 1.18. Suppose the joint density function of X and Y is given by $f(x,y) = 1/2x \exp(-xy)$ where $x \in [3, 5]$ and $y \in (0, \infty)$.

1. Give the form of $E[Y|X=x]$.
2. Graph the conditional expectation curve.

Exercise 1.19. Suppose a fund manager has \$10,000 of yours under management and tells you that the expected value of your portfolio in two years time is \$30,000 and that with probability 75% your investment will be worth at least \$40,000 in two years time.

1. Do you believe her?

2. Next, suppose she tells you that the standard deviation of your portfolio value is 2,000. Assuming this is true (as is the expected value), what is the most you can say about the probability your portfolio value falls between \$20,000 and \$40,000 in two years time?

Exercise 1.20. Suppose the joint probability density function of two random variables is given by $f(x) = \frac{2}{5}(3x + 2y)$ where $x \in [0, 1]$ and $y \in [0, 1]$.

1. What is the marginal probability density function of X ?
2. What is $E[X|Y = y]$? Are X and Y independent? (Hint: What must the form of $E[X|Y]$ be when they are independent?)

Exercise 1.21. Let Y be distributed χ^2_{15} .

1. What is $\Pr(y > 27.488)$?
2. What is $\Pr(6.262 \leq y \leq 27.488)$?
3. Find C where $\Pr(y \geq c) = \alpha$ for $\alpha \in \{0.01, 0.05, 0.01\}$.
Next, Suppose Z is distributed χ^2_5 and is independent of Y .
4. Find C where $\Pr(y + z \geq c) = \alpha$ for $\alpha \in \{0.01, 0.05, 0.01\}$.

Exercise 1.22. Suppose X is a bivariate random variable with parameters

$$\mu = \begin{bmatrix} 5 \\ 8 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix}.$$

1. What is $E[X_1|X_2]$?
2. What is $V[X_1|X_2]$?
3. Show (numerically) that the law of total variance holds for X_2 .

Exercise 1.23. Suppose $y \sim N(5, 36)$ and $x \sim N(4, 25)$ where X and Y are independent.

1. What is $\Pr(y > 10)$?
2. What is $\Pr(-10 < y < 10)$?
3. What is $\Pr(x - y > 0)$?
4. Find C where $\Pr(x - y > C) = \alpha$ for $\alpha \in \{0.10, 0.05, 0.01\}$?

Rectangle Method				
Bins	$\pm 3\sigma$	$\pm 4\sigma$	$\pm 6\sigma$	$\pm 10\sigma$
10	7.19	7.43	7.58	8.50
20	7.13	7.35	7.39	7.50
50	7.12	7.33	7.34	7.36
1000	7.11	7.32	7.33	7.33

Trapezoid Method				
Bins	$\pm 3\sigma$	$\pm 4\sigma$	$\pm 6\sigma$	$\pm 10\sigma$
10	6.96	7.11	6.86	5.53
20	7.08	7.27	7.22	7.01
50	7.11	7.31	7.31	7.28
1000	7.11	7.32	7.33	7.33

Simpson's Rule				
Bins	$\pm 3\sigma$	$\pm 4\sigma$	$\pm 6\sigma$	$\pm 10\sigma$
10	7.11	7.32	7.34	7.51
20	7.11	7.32	7.33	7.34
50	7.11	7.32	7.33	7.33
1000	7.11	7.32	7.33	7.33

Monte Carlo				
Draws (m)	100	1000		
Mean	7.34	7.33		
Std. Dev.	0.88	0.28		

Table 1.1: Computed values for the expected payout for an option, where the correct value is 7.33. The top three panels use approximations to the function which have simple to compute areas. The bottom panel shows the average and standard deviation from a Monte Carlo integration where the number of points varies and 10,000 simulations were used.

Chapter 2

Estimation, Inference, and Hypothesis Testing

Note: The primary reference for these notes is Ch. 7 and 8 of Casella and Berger (2001). This text may be challenging if new to this topic and Ch. 7 – 10 of Wackerly, Mendenhall, and Scheaffer (2001) may be useful as an introduction.

This chapter provides an overview of estimation, distribution theory, inference, and hypothesis testing. Testing an economic or financial theory is a multi-step process. First, any unknown parameters must be estimated. Next, the distribution of the estimator must be determined. Finally, formal hypothesis tests must be conducted to examine whether the data are consistent with the theory. This chapter is intentionally “generic” by design and focuses on the case where the data are independent and identically distributed. Properties of specific models will be studied in detail in the chapters on linear regression, time series, and univariate volatility modeling.

Three steps must be completed to test the implications of an economic theory:

- Estimate unknown parameters
- Determine the distributional of estimator
- Conduct hypothesis tests to examine whether the data are compatible with a theoretical model

This chapter covers each of these steps with a focus on the case where the data is independent and identically distributed (i.i.d.). The heterogeneous but independent case will be covered in the chapter on linear regression and the dependent case will be covered in the chapters on time series.

2.1 Estimation

Once a model has been specified and hypotheses postulated, the first step is to estimate the parameters of the model. Many methods are available to accomplish this task. These include parametric, semi-parametric, semi-nonparametric and nonparametric estimators and a variety of estimation methods often classified as M-, R- and L-estimators.¹

¹There is another important dimension in the categorization of estimators: Bayesian or frequentist. Bayesian estimators make use of Bayes rule to perform inference about unknown quantities – parameters – conditioning on the observed

Parametric models are tightly parameterized and have desirable statistical properties when their specification is correct, such as providing consistent estimates with small variances. Nonparametric estimators are more flexible and avoid making strong assumptions about the relationship between variables. This allows nonparametric estimators to capture a wide range of relationships but comes at the cost of precision. In many cases, nonparametric estimators are said to have a *slower rate of convergence* than similar parametric estimators. The practical consequence of the rate is that nonparametric estimators are desirable when there is a proliferation of data and the relationships between variables may be difficult to postulate *a priori*. In situations where less data is available, or when an economic model proffers a relationship among variables, parametric estimators are generally preferable.

Semi-parametric and semi-nonparametric estimators bridge the gap between fully parametric estimators and nonparametric estimators. Their difference lies in “how parametric” the model and estimator are. Estimators which postulate parametric relationships between variables but estimate the underlying distribution of errors flexibly are known as semi-parametric. Estimators which take a stand on the distribution of the errors but allow for flexible relationships between variables are semi-nonparametric. This chapter focuses exclusively on parametric models and estimators. This choice is more reflective of the common practice than a critique of nonparametric methods.

Another important characterization of estimators is whether they are members of the M-, L- or R-estimator classes.² M-estimators (also known as extremum estimators) always involve maximizing or minimizing some objective function. M-estimators are the most commonly used class in financial econometrics and include maximum likelihood, regression, classical minimum distance and both the classical and the generalized method of moments. L-estimators, also known as linear estimators, are a class where the estimator can be expressed as a linear function of ordered data. Members of this family can always be written as

$$\sum_{i=1}^n w_i y_i$$

for some set of weights $\{w_i\}$ where the data, y_i , are ordered such that $y_{j-1} \leq y_j$ for $j = 2, 3, \dots, n$. This class of estimators obviously includes the sample mean by setting $w_i = \frac{1}{n}$ for all i , and also includes the median by setting $w_i = 0$ for all i except $w_j = 1$ where $j = (n+1)/2$ (n is odd) or $w_j = w_{j+1} = 1/2$ where $j = n/2$ (n is even). R-estimators exploit the *rank* of the data. Common examples of R-estimators include the minimum, maximum and Spearman’s rank correlation, which is the usual correlation estimator on the ranks of the data rather than on the data themselves. Rank statistics are often robust to outliers and non-linearities.

2.1.1 M-Estimators

The use of M-estimators is pervasive in financial econometrics. Three common types of M-estimators include the method of moments, both classical and generalized, maximum likelihood and classical minimum distance.

data. Frequentist estimators rely on randomness averaging out across observations. Frequentist methods are dominant in financial econometrics although the use of Bayesian methods has been recently increasing.

²Many estimators are members of more than one class. For example, the median is a member of all three.

2.1.2 Maximum Likelihood

Maximum likelihood uses the distribution of the data to estimate any unknown parameters by finding the values which make the data as likely as possible to have been observed – in other words, by maximizing the likelihood. Maximum likelihood estimation begins by specifying the *joint* distribution, $f(\mathbf{y}; \theta)$, of the observable data, $\mathbf{y} = \{y_1, y_2, \dots, y_n\}$, as a function of a k by 1 vector θ which contains all parameters. Note that this is the joint density, and so it includes both the information in the marginal distributions of y_i and information relating the marginals to one another.³ Maximum likelihood estimation “reverses” the likelihood to express the probability of θ in terms of the observed \mathbf{y} , $L(\theta; \mathbf{y}) = f(\mathbf{y}; \theta)$.

The maximum likelihood estimator, $\hat{\theta}$, is defined as the solution to

$$\hat{\theta} = \arg \max_{\theta} L(\theta; \mathbf{y}) \quad (2.1)$$

where $\arg \max$ is used in place of \max to indicate that the maximum may not be unique – it could be set valued – and to indicate that the *global* maximum is required.⁴ Since $L(\theta; \mathbf{y})$ is strictly positive, the log of the likelihood can be used to estimate θ .⁵ The log-likelihood is defined as $l(\theta; \mathbf{y}) = \ln L(\theta; \mathbf{y})$. In most situations the maximum likelihood estimator (MLE) can be found by solving the k by 1 score vector,

$$\frac{\partial l(\theta; \mathbf{y})}{\partial \theta} = \mathbf{0}$$

although a score-based solution does not work when θ is constrained and $\hat{\theta}$ lies on the boundary of the parameter space or when the permissible range of values for y_i depends on θ . The first problem is common enough that it is worth keeping in mind. It is particularly common when working with variances which must be (weakly) positive by construction. The second issue is fairly rare in financial econometrics.

2.1.2.1 Maximum Likelihood Estimation of a Poisson Model

Realizations from a Poisson process are non-negative and discrete. The Poisson is common in ultra-high-frequency econometrics where the usual assumption that prices lie in a continuous space is

³Formally the relationship between the marginal is known as the *copula*. Copulas and their use in financial econometrics will be explored in the second term.

⁴Many likelihoods have more than one maximum (i.e. local maxima). The maximum likelihood estimator is always defined as the global maximum.

⁵Note that the log transformation is strictly increasing and globally concave. If z^* is the maximum of $g(z)$, and thus

$$\left. \frac{\partial g(z)}{\partial z} \right|_{z=z^*} = 0$$

then z^* must also be the maximum of $\ln(g(z))$ since

$$\left. \frac{\partial \ln(g(z))}{\partial z} \right|_{z=z^*} = \left. \frac{g'(z)}{g(z)} \right|_{z=z^*} = \frac{0}{g(z^*)} = 0$$

which follows since $g(z) > 0$ for any value of z .

implausible. For example, trade prices of US equities evolve on a grid of prices typically separated by \$0.01. Suppose $y_i \stackrel{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$. The pdf of a single observation is

$$f(y_i; \lambda) = \frac{\exp(-\lambda)\lambda^{y_i}}{y_i!} \quad (2.2)$$

and since the data are independent and identically distributed (i.i.d.), the joint likelihood is simply the product of the n individual likelihoods,

$$f(\mathbf{y}; \lambda) = L(\lambda; \mathbf{y}) = \prod_{i=1}^n \frac{\exp(-\lambda)\lambda^{y_i}}{y_i!}.$$

The log-likelihood is

$$l(\lambda; \mathbf{y}) = \sum_{i=1}^n -\lambda + y_i \ln(\lambda) - \ln(y_i!) \quad (2.3)$$

which can be further simplified to

$$l(\lambda; \mathbf{y}) = -n\lambda + \ln(\lambda) \sum_{i=1}^n y_i - \sum_{j=1}^n \ln(y_j!)$$

The first derivative is

$$\frac{\partial l(\lambda; \mathbf{y})}{\partial \lambda} = -n + \lambda^{-1} \sum_{i=1}^n y_i. \quad (2.4)$$

The MLE is found by setting the derivative to 0 and solving,

$$\begin{aligned} -n + \hat{\lambda}^{-1} \sum_{i=1}^n y_i &= 0 \\ \hat{\lambda}^{-1} \sum_{i=1}^n y_i &= n \\ \sum_{i=1}^n y_i &= n\hat{\lambda} \\ \hat{\lambda} &= n^{-1} \sum_{i=1}^n y_i \end{aligned}$$

Thus the maximum likelihood estimator in a Poisson is the sample mean.

2.1.2.2 Maximum Likelihood Estimation of a Normal (Gaussian) Model

Suppose y_i is assumed to be i.i.d. normally distributed with mean μ and variance σ^2 . The pdf of a normal is

$$f(y_i; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2}\right). \quad (2.5)$$

where $\theta = [\mu \ \sigma^2]'$. The joint likelihood is the product of the n individual likelihoods,

$$f(\mathbf{y}; \theta) = L(\theta; \mathbf{y}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2}\right).$$

Taking logs,

$$l(\theta; \mathbf{y}) = \sum_{i=1}^n -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\sigma^2) - \frac{(y_i - \mu)^2}{2\sigma^2} \quad (2.6)$$

which can be simplified to

$$l(\theta; \mathbf{y}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2} \sum_{i=1}^n \frac{(y_i - \mu)^2}{\sigma^2}.$$

Taking the derivative with respect to the parameters $\theta = (\mu, \sigma^2)'$,

$$\frac{\partial l(\theta; \mathbf{y})}{\partial \mu} = \sum_{i=1}^n \frac{(y_i - \mu)}{\sigma^2} \quad (2.7)$$

$$\frac{\partial l(\theta; \mathbf{y})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2} \sum_{i=1}^n \frac{(y_i - \mu)^2}{\sigma^4}. \quad (2.8)$$

Setting these equal to zero, the first condition can be directly solved by multiplying both sides by $\hat{\sigma}^2$, assumed positive, and the estimator for μ is the sample average.

$$\begin{aligned} \sum_{i=1}^n \frac{(y_i - \hat{\mu})}{\hat{\sigma}^2} &= 0 \\ \hat{\sigma}^2 \sum_{i=1}^n \frac{(y_i - \hat{\mu})}{\hat{\sigma}^2} &= \hat{\sigma}^2 0 \\ \sum_{i=1}^n y_i - n\hat{\mu} &= 0 \\ n\hat{\mu} &= \sum_{i=1}^n y_i \\ \hat{\mu} &= n^{-1} \sum_{i=1}^n y_i \end{aligned}$$

Plugging this value into the second score and setting equal to 0, the ML estimator of σ^2 is

$$\begin{aligned} -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2} \sum_{i=1}^n \frac{(y_i - \hat{\mu})^2}{\hat{\sigma}^4} &= 0 \\ 2\hat{\sigma}^4 \left(-\frac{n}{2\hat{\sigma}^2} + \frac{1}{2} \sum_{i=1}^n \frac{(y_i - \hat{\mu})^2}{\hat{\sigma}^4} \right) &= 2\hat{\sigma}^4 0 \\ -n\hat{\sigma}^2 + \sum_{i=1}^n (y_i - \hat{\mu})^2 &= 0 \\ \hat{\sigma}^2 &= n^{-1} \sum_{i=1}^n (y_i - \hat{\mu})^2 \end{aligned}$$

2.1.3 Conditional Maximum Likelihood

Interest often lies in the distribution of a random variable conditional on one or more observed values, where the distribution of the observed values is not of interest. When this occurs, it is natural to use conditional maximum likelihood. Suppose interest lies in modeling a random variable Y conditional on one or more variables \mathbf{X} . The likelihood for a single observation is $f_i(y_i|\mathbf{x}_i)$, and when Y_i are conditionally i.i.d., then

$$L(\theta; \mathbf{y}|\mathbf{X}) = \prod_{i=1}^n f(y_i|\mathbf{x}_i),$$

and the log-likelihood is

$$l(\theta; \mathbf{y}|\mathbf{X}) = \sum_{i=1}^n \ln f(y_i|\mathbf{x}_i).$$

The conditional likelihood is not usually sufficient to estimate parameters since the relationship between Y and \mathbf{X} has not been specified. Conditional maximum likelihood specifies the model parameters conditionally on \mathbf{x}_i . For example, in an conditional normal, $y|\mathbf{x}_i \sim N(\mu_i, \sigma^2)$ where $\mu_i = g(\beta, \mathbf{x}_i)$ is some function which links parameters and conditioning variables. In many applications a linear relationship is assumed so that

$$\begin{aligned} y_i &= \beta' \mathbf{x}_i + \varepsilon_i \\ &= \sum_{j=1}^k \beta_j x_{i,j} + \varepsilon_i \\ &= \mu_i + \varepsilon_i. \end{aligned}$$

Other relationships are possible, including functions $g(\beta' \mathbf{x}_i)$ which limits to range of $\beta' \mathbf{x}_i$ such as $\exp(\beta' \mathbf{x}_i)$ (positive numbers), the normal cdf ($\Phi(\beta' \mathbf{x})$) or the logistic function,

$$\Lambda(\beta' \mathbf{x}_i) = \exp(\beta' \mathbf{x}_i) / (1 + \exp(\beta' \mathbf{x}_i)),$$

since both limit the range to $(0, 1)$.

2.1.3.1 Example: Conditional Bernoulli

Suppose Y_i and X_i are Bernoulli random variables where the conditional distribution of Y_i given X_i is

$$y_i|x_i \sim \text{Bernoulli}(\theta_0 + \theta_1 x_i)$$

so that the conditional probability of observing a success ($y_i = 1$) is $p_i = \theta_0 + \theta_1 x_i$. The conditional likelihood is

$$L(\theta; \mathbf{y}|\mathbf{x}) = \prod_{i=1}^n (\theta_0 + \theta_1 x_i)^{y_i} (1 - (\theta_0 + \theta_1 x_i))^{1-y_i},$$

the conditional log-likelihood is

$$l(\theta; \mathbf{y}|\mathbf{x}) = \sum_{i=1}^n y_i \ln(\theta_0 + \theta_1 x_i) + (1 - y_i) \ln(1 - (\theta_0 + \theta_1 x_i)),$$

and the maximum likelihood estimator can be found by differentiation.

$$\begin{aligned} \frac{\partial l(\hat{\theta}; \mathbf{y}|\mathbf{x})}{\partial \hat{\theta}_0} &= \sum_{i=1}^n \frac{y_i}{\hat{\theta}_0 + \hat{\theta}_1 x_i} - \frac{1 - y_i}{1 - \hat{\theta}_0 - \hat{\theta}_1 x_i} = 0 \\ \frac{\partial l(\theta; \mathbf{y}|\mathbf{x})}{\partial \theta_1} &= \sum_{i=1}^n \frac{x_i y_i}{\hat{\theta}_0 + \hat{\theta}_1 x_i} - \frac{x_i (1 - y_i)}{1 - \hat{\theta}_0 - \hat{\theta}_1 x_i} = 0. \end{aligned}$$

Using the fact that x_i is also Bernoulli, the second score can be solved

$$\begin{aligned} 0 = \sum_{i=1}^n x_i \left(\frac{y_i}{\hat{\theta}_0 + \hat{\theta}_1} + \frac{(1 - y_i)}{(1 - \hat{\theta}_0 - \hat{\theta}_1)} \right) &= \sum_{i=1}^n \frac{n_{xy}}{\hat{\theta}_0 + \hat{\theta}_1} - \frac{n_x - n_{xy}}{1 - \hat{\theta}_0 - \hat{\theta}_1 x_i} \\ &= n_{xy} (1 - (\hat{\theta}_0 + \hat{\theta}_1)) - (n_x - n_{xy}) (\hat{\theta}_0 + \hat{\theta}_1) \\ &= n_{xy} - n_{xy} (\hat{\theta}_0 + \hat{\theta}_1) - n_x (\hat{\theta}_0 + \hat{\theta}_1) + n_{xy} (\hat{\theta}_0 + \hat{\theta}_1) \\ \hat{\theta}_0 + \hat{\theta}_1 &= \frac{n_{xy}}{n_x}, \end{aligned}$$

Define $n_x = \sum_{i=1}^n x_i$, $n_y = \sum_{i=1}^n y_i$ and $n_{xy} = \sum_{i=1}^n x_i y_i$. The first score than also be rewritten as

$$\begin{aligned} 0 = \sum_{i=1}^n \frac{y_i}{\hat{\theta}_0 + \hat{\theta}_1 x_i} - \frac{1 - y_i}{1 - \hat{\theta}_0 - \hat{\theta}_1 x_i} &= \sum_{i=1}^n \frac{y_i (1 - x_i)}{\hat{\theta}_0} + \frac{y_i x_i}{\hat{\theta}_0 + \hat{\theta}_1} - \frac{1 - y_i (1 - x_i)}{1 - \hat{\theta}_0} - \frac{(1 - y_i) x_i}{1 - \hat{\theta}_0 - \hat{\theta}_1} \\ &= \sum_{i=1}^n \frac{y_i (1 - x_i)}{\hat{\theta}_0} - \frac{1 - y_i (1 - x_i)}{1 - \hat{\theta}_0} + \left\{ \frac{x_i y_i}{\hat{\theta}_0 + \hat{\theta}_1} - \frac{x_i (1 - y_i)}{1 - \hat{\theta}_0 - \hat{\theta}_1} \right\} \\ &= \frac{n_y - n_{xy}}{\hat{\theta}_0} - \frac{n - n_y - n_x + n_{xy}}{1 - \hat{\theta}_0} + \{0\} \\ &= n_y - n_{xy} - \hat{\theta}_0 n_y + \hat{\theta}_0 n - \hat{\theta}_0 n + \hat{\theta}_0 n_y + \hat{\theta}_0 n_x - \hat{\theta}_0 n_{xy} \\ \hat{\theta}_0 &= \frac{n_y - n_{xy}}{n - n_x} \end{aligned}$$

so that $\hat{\theta}_1 = \frac{n_{xy}}{n_x} - \frac{n_y - n_{xy}}{n - n_x}$. The “0” in the previous derivation follows from noting that the quantity in {} is equivalent to the first score and so is 0 at the MLE. If X_i was not a Bernoulli random variable, then it would not be possible to analytically solve this problem. In these cases, numerical methods are needed.⁶

2.1.3.2 Example: Conditional Normal

Suppose $\mu_i = \beta x_i$ where Y_i given X_i is conditionally normal. Assuming that Y_i are conditionally i.i.d., the likelihood and log-likelihood are

$$\begin{aligned} L(\theta; \mathbf{y} | \mathbf{x}) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \beta x_i)^2}{2\sigma^2}\right) \\ l(\theta; \mathbf{y} | \mathbf{x}) &= \sum_{i=1}^n -\frac{1}{2} \left(\ln(2\pi) + \ln(\sigma^2) + \frac{(y_i - \beta x_i)^2}{\sigma^2} \right). \end{aligned}$$

The scores of the likelihood are

$$\begin{aligned} \frac{\partial l(\theta; \mathbf{y} | \mathbf{x})}{\partial \beta} &= \sum_{i=1}^n \frac{x_i (y_i - \hat{\beta} x_i)}{\hat{\sigma}^2} = 0 \\ \frac{\partial l(\theta; \mathbf{y} | \mathbf{x})}{\partial \sigma^2} &= -\frac{1}{2} \sum_{i=1}^n \frac{1}{\hat{\sigma}^2} - \frac{(y_i - \hat{\beta} x_i)^2}{(\hat{\sigma}^2)^2} = 0 \end{aligned}$$

After multiplying both sides the first score by $\hat{\sigma}^2$, and both sides of the second score by $-2\hat{\sigma}^4$, solving the scores is straight forward, and so

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{i=1}^n x_i y_i}{\sum_{j=1}^n x_j^2} \\ \hat{\sigma}^2 &= n^{-1} \sum_{i=1}^n (y_i - \hat{\beta} x_i)^2. \end{aligned}$$

2.1.3.3 Example: Conditional Poisson

Suppose Y_i is conditional on X_1 i.i.d. distributed $\text{Poisson}(\lambda_i)$ where $\lambda_i = \exp(\theta x_i)$. The likelihood and log-likelihood are

$$\begin{aligned} L(\theta; \mathbf{y} | \mathbf{x}) &= \prod_{i=1}^n \frac{\exp(-\lambda_i) \lambda_i^{y_i}}{y_i!} \\ l(\theta; \mathbf{y} | \mathbf{x}) &= \sum_{i=1}^n \exp(\theta x_i) + y_i (\theta x_i) - \ln(y_i!). \end{aligned}$$

⁶When X_i is not Bernoulli, it is also usually necessary to use a function to ensure p_i , the conditional probability, is in $[0, 1]$. Two common choices are the normal cdf and the logistic function.

The score of the likelihood is

$$\frac{\partial l(\theta; \mathbf{y} | \mathbf{x})}{\partial \theta} = \sum_{i=1}^n -x_i \exp(\hat{\theta} x_i) + x_i y_i = 0.$$

This score cannot be analytically solved and so a numerical optimizer must be used to find the solution. It is possible, however, to show the score has conditional expectation 0 since $E[Y_i | X_i] = \lambda_i$.

$$\begin{aligned} E\left[\frac{\partial l(\theta; \mathbf{y} | \mathbf{x})}{\partial \theta} | \mathbf{X}\right] &= E\left[\sum_{i=1}^n -x_i \exp(\theta x_i) + x_i y_i | \mathbf{X}\right] \\ &= \sum_{i=1}^n E[-x_i \exp(\theta x_i) | \mathbf{X}] + E[x_i y_i | \mathbf{X}] \\ &= \sum_{i=1}^n -x_i \lambda_i + x_i E[y_i | \mathbf{X}] \\ &= \sum_{i=1}^n -x_i \lambda_i + x_i \lambda_i = 0. \end{aligned}$$

2.1.4 The Method of Moments

The Method of moments, often referred to as the classical method of moments to differentiate it from the *generalized* method of moments (GMM, chapter 6) uses the data to match *noncentral* moments.

Definition 2.1 (Noncentral Moment). The r^{th} noncentral moment is defined

$$\mu'_r \equiv E[X^r] \tag{2.9}$$

for $r = 1, 2, \dots$

Central moments are similarly defined, only centered around the mean.

Definition 2.2 (Central Moment). The r^{th} central moment is defined

$$\mu_r \equiv E[(X - \mu'_1)^r] \tag{2.10}$$

for $r = 2, 3, \dots$ where the 1st central moment is defined to be equal to the 1st noncentral moment.

Since $E[x_i^r]$ is not known any estimator based on it is *infeasible*. The obvious solution is to use the *sample analogue* to estimate its value, and the *feasible* method of moments estimator is

$$\hat{\mu}'_r = n^{-1} \sum_{i=1}^n x_i^r, \tag{2.11}$$

the sample average of the data raised to the r^{th} power. While the classical method of moments was originally specified using noncentral moments, the central moments are usually the quantities of interest. The central moments can be directly estimated,

$$\hat{\mu}_r = n^{-1} \sum_{i=1}^n (x_i - \hat{\mu}_1)^r, \quad (2.12)$$

and so can be simply implemented by first estimating the mean ($\hat{\mu}_1$) and then estimating the remaining central moments. An alternative is to expand the noncentral moment in terms of central moments. For example, the second noncentral moment can be expanded in terms of the first two central moments,

$$\mu'_2 = \mu_2 + \mu_1^2$$

which is the usual identity that states that expectation of a random variable squared, $E[x_i^2]$, is equal to the variance, $\mu_2 = \sigma^2$, plus the mean squared, μ_1^2 . Likewise, it is easy to show that

$$\mu'_3 = \mu_3 + 3\mu_2\mu_1 + \mu_1^3$$

directly by expanding $E[(X - \mu_1)^3]$ and solving for μ'_3 . To understand that the method of moments is in the class of M-estimators, note that the expression in eq. (2.12) is the first order condition of a simple quadratic form,

$$\arg \min_{\mu, \mu_2, \dots, \mu_k} \left(n^{-1} \sum_{i=1}^n x_i - \mu_1 \right)^2 + \sum_{j=2}^k \left(n^{-1} \sum_{i=1}^n (x_i - \mu)^j - \mu_j \right)^2, \quad (2.13)$$

and since the number of unknown parameters is identical to the number of equations, the solution is exact.⁷

2.1.4.1 Method of Moments Estimation of the Mean and Variance

The classical method of moments estimator for the mean and variance for a set of i.i.d. data $\{y_i\}_{i=1}^n$ where $E[Y_i] = \mu$ and $E[(Y_i - \mu)^2] = \sigma^2$ is given by estimating the first two noncentral moments and then solving for σ^2 .

$$\begin{aligned} \hat{\mu} &= n^{-1} \sum_{i=1}^n y_i \\ \hat{\sigma}^2 + \hat{\mu}^2 &= n^{-1} \sum_{i=1}^n y_i^2 \end{aligned}$$

and thus the variance estimator is $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n y_i^2 - \hat{\mu}^2$. Following some algebra, it is simple to show that the central moment estimator could be used equivalently, and so $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (y_i - \hat{\mu})^2$.

⁷Note that μ_1 , the mean, is generally denoted with the subscript suppressed as μ .

2.1.4.2 Method of Moments Estimation of the Range of a Uniform

Consider a set of realization of a random variable with a uniform density over $[0, \theta]$, and so $y_i \stackrel{\text{i.i.d.}}{\sim} U(0, \theta)$. The expectation of y_i is $E[Y_i] = \theta/2$, and so the method of moments estimator for the upper bound is

$$\hat{\theta} = 2n^{-1} \sum_{i=1}^n y_i.$$

2.1.5 Classical Minimum Distance

A third – and less frequently encountered – type of M-estimator is classical minimum distance (CMD) which is also known as minimum χ^2 in some circumstances. CMD differs from MLE and the method of moments in that it is an estimator that operates using initial parameter estimates produced by another estimator rather than on the data directly. CMD is most common when a simple MLE or moment-based estimator is available that can estimate a model without some economically motivated constraints on the parameters. This initial estimator, $\hat{\psi}$ is then used to estimate the parameters of the model, θ , by minimizing a quadratic function of the form

$$\hat{\theta} = \arg \min_{\theta} (\hat{\psi} - \mathbf{g}(\theta))' \mathbf{W} (\hat{\psi} - \mathbf{g}(\theta)) \quad (2.14)$$

where \mathbf{W} is a positive definite weighting matrix. When \mathbf{W} is chosen as the covariance of $\hat{\psi}$, the CMD estimator becomes the minimum- χ^2 estimator since outer products of standardized normals are χ^2 random variables.

2.2 Convergence and Limits for Random Variables

Before turning to properties of estimators, it is useful to discuss some common measures of convergence for sequences. Before turning to the alternative definitions which are appropriate for random variables, recall the definition of a limit of a non-stochastic sequence.

Definition 2.3 (Limit). Let $\{x_n\}$ be a non-stochastic sequence. If there exists N such that for every $n > N$, $|x_n - x| < \epsilon \forall \epsilon > 0$, when x is called the limit of x_n . When this occurs, $x_n \rightarrow x$ or $\lim_{n \rightarrow \infty} x_n = x$.

A limit is a point where a sequence will approach, and eventually, always remain near. It isn't necessary that the limit is ever attained, only that for any choice of $\epsilon > 0$, x_n will eventually always be less than ϵ away from its limit.

Limits of random variables come in many forms. The first the type of convergence is both the weakest and most abstract.

Definition 2.4 (Convergence in Distribution). Let $\{\mathbf{Y}_n\}$ be a sequence of random variables and let $\{F_n\}$ be the associated sequence of cdfs. If there exists a cdf F where $F_n(\mathbf{y}) \rightarrow F(\mathbf{y})$ for all \mathbf{y} where F is continuous, then F is the limiting cdf of $\{\mathbf{Y}_n\}$. Let \mathbf{Y} be a random variable with cdf F , then \mathbf{Y}_n converges in distribution to $\mathbf{Y} \sim F$, $\mathbf{Y}_n \xrightarrow{d} \mathbf{Y} \sim F$, or simply $\mathbf{Y}_n \xrightarrow{d} F$.

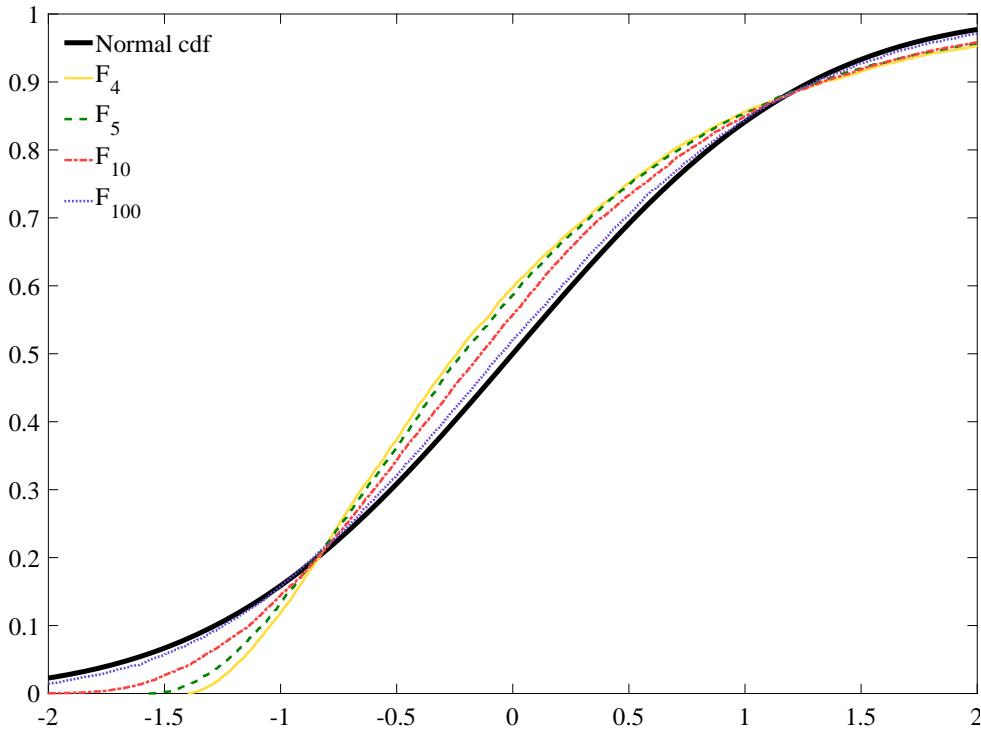


Figure 2.1: This figure shows a sequence of cdfs $\{F_i\}$ that converge to the cdf of a standard normal.

Convergence in distribution means that the limiting cdf of a sequence of random variables is the same as the convergent random variable. This is a very weak form of convergence since all it requires is that the distributions are the same. For example, suppose $\{X_n\}$ is an i.i.d. sequence of standard normal random variables, and Y is a standard normal random variable. X_n trivially converges to distribution to Y ($X_n \xrightarrow{d} Y$) even though Y is completely independent of $\{X_n\}$ – the limiting cdf of X_n is merely the same as the cdf of Y . Despite the weakness of convergence in distribution, it is an essential notion of convergence that is used to perform inference on estimated parameters.

Figure 2.1 shows an example of a sequence of random variables which converge in distribution. The sequence is

$$X_n = \sqrt{n} \frac{\sum_{i=1}^n Y_i - 1}{\sqrt{2}}$$

where Y_i are i.i.d. χ_1^2 random variables. This is a studentized average since the variance of the average is $2/n$ and the mean is 1. By the time $n = 100$, F_{100} is nearly indistinguishable from the standard normal cdf.

Convergence in distribution is preserved through functions.

Theorem 2.1 (Continuous Mapping Theorem). *Let $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ and let the random variable $g(\mathbf{X})$ be defined by a function $g(\mathbf{x})$ that is continuous everywhere except possibly on a set with zero probability. Then $g(\mathbf{X}_n) \xrightarrow{d} g(\mathbf{X})$.*

The continuous mapping theorem is useful since it facilitates the study of functions of sequences of random variables. For example, in hypothesis testing, it is common to use quadratic forms of nor-

mals, and when appropriately standardized, quadratic forms of normally distributed random variables follow a χ^2 distribution.

The next form of convergence is stronger than convergence in distribution since the limit is to a specific target, not just a cdf.

Definition 2.5 (Convergence in Probability). The sequence of random variables $\{\mathbf{X}_n\}$ converges in probability to \mathbf{X} if and only if

$$\lim_{n \rightarrow \infty} \Pr(|X_{i,n} - X_i| < \varepsilon) = 1 \quad \forall \varepsilon > 0, \forall i.$$

When this holds, $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$ or equivalently $\text{plim } \mathbf{X}_n = \mathbf{X}$ (or $\text{plim } \mathbf{X}_n - \mathbf{X} = 0$) where plim is probability limit.

Note that \mathbf{X} can be either a random variable or a constant (degenerate random variable). For example, if $X_n = n^{-1} + Z$ where Z is a normally distributed random variable, then $X_n \xrightarrow{P} Z$. Convergence in probability requires virtually all of the probability mass of \mathbf{X}_n to lie near \mathbf{X} . This is a very weak form of convergence since it is possible that a small amount of probability can be arbitrarily far away from \mathbf{X} . Suppose a scalar random sequence $\{X_n\}$ takes the value 0 with probability $1 - 1/n$ and n with probability $1/n$. Then $\{X_n\} \xrightarrow{P} 0$ although $E[X_n] = 1$ for all n .

Convergence in probability, however, is strong enough that it is useful work studying random variables and functions of random variables.

Theorem 2.2. Let $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$ and let the random variable $g(\mathbf{X})$ be defined by a function $g(x)$ that is continuous everywhere except possibly on a set with zero probability. Then $g(\mathbf{X}_n) \xrightarrow{P} g(\mathbf{X})$ (or equivalently $\text{plim } g(\mathbf{X}_n) = g(\mathbf{X})$).

This theorem has some, simple useful forms. Suppose the k -dimensional vector $\mathbf{X}_n \xrightarrow{P} \mathbf{X}$, the conformable vector $\mathbf{Y}_n \xrightarrow{P} \mathbf{Y}$ and \mathbf{C} is a conformable constant matrix, then

- $\text{plim } \mathbf{C}\mathbf{X}_n = \mathbf{C}\mathbf{X}$
- $\text{plim } \sum_{i=1}^k X_{i,n} = \sum_{i=1}^k \text{plim } X_{i,n}$ – the plim of the sum is the sum of the plims
- $\text{plim } \prod_{i=1}^k X_{i,n} = \prod_{i=1}^k \text{plim } X_{i,n}$ – the plim of the product is the product of the plims
- $\text{plim } \mathbf{Y}_n \mathbf{X}_n = \mathbf{Y}\mathbf{X}$
- When \mathbf{Y}_n is a square matrix and \mathbf{Y} is nonsingular, then $\mathbf{Y}_n^{-1} \xrightarrow{P} \mathbf{Y}^{-1}$ – the inverse function is continuous and so plim of the inverse is the inverse of the plim
- When \mathbf{Y}_n is a square matrix and \mathbf{Y} is nonsingular, then $\mathbf{Y}_n^{-1} \mathbf{X}_n \xrightarrow{P} \mathbf{Y}^{-1} \mathbf{X}$.

These properties are very difference from the expectations operator. In particular, the plim operator passes through functions which allows for broad application. For example,

$$E\left[\frac{1}{X}\right] \neq \frac{1}{E[X]}$$

whenever X is a non-degenerate random variable. However, if $X_n \xrightarrow{p} X$, then

$$\begin{aligned}\text{plim} \frac{1}{X_n} &= \frac{1}{\text{plim} X_n} \\ &= \frac{1}{X}.\end{aligned}$$

Alternative definitions of convergence strengthen convergence in probability. In particular, convergence in mean square requires that the expected squared deviation must be zero. This requires that $E[X_n] = X$ and $V[X_n] = 0$.

Definition 2.6 (Convergence in Mean Square). The sequence of random variables $\{\mathbf{X}_n\}$ converges in mean square to \mathbf{X} if and only if

$$\lim_{n \rightarrow \infty} E[(X_{i,n} - X_i)^2] = 0, \forall i.$$

When this holds, $\mathbf{X}_n \xrightarrow{m.s.} \mathbf{X}$.

Mean square convergence is strong enough to ensure that, when the limit is random \mathbf{X} than $E[\mathbf{X}_n] = E[\mathbf{X}]$ and $V[\mathbf{X}_n] = V[\mathbf{X}]$ – these relationships do not necessarily hold when only $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$.

Theorem 2.3 (Convergence in mean square implies consistency). *If $\mathbf{X}_n \xrightarrow{m.s.} \mathbf{X}$ then $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$.*

This result follows directly from Chebyshev's inequality. A final, and very strong, measure of convergence for random variables is known as almost sure convergence.

Definition 2.7 (Almost sure convergence). The sequence of random variables $\{\mathbf{X}_n\}$ converges almost surely to \mathbf{X} if and only if

$$\lim_{n \rightarrow \infty} \Pr(X_{i,n} - X_i = 0) = 1, \forall i.$$

When this holds, $\mathbf{X}_n \xrightarrow{a.s.} \mathbf{X}$.

Almost sure convergence requires all probability to be on the limit point. This is a stronger condition than either convergence in probability or convergence in mean square, both of which allow for some probability to be (relatively) far from the limit point.

Theorem 2.4 (Almost sure convergence implications). *If $\mathbf{X}_n \xrightarrow{a.s.} \mathbf{X}$ then $\mathbf{X}_n \xrightarrow{m.s.} \mathbf{X}$ and $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$.*

Random variables which converge almost surely to a limit are asymptotically degenerate on that limit.

The Slutsky theorem combines variables which converge in distribution with variables which converge in probability to show that the joint limit of functions behaves as expected.

Theorem 2.5 (Slutsky Theorem). *Let $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ and let $\mathbf{Y} \xrightarrow{p} \mathbf{C}$, a constant, then for conformable \mathbf{X} and \mathbf{C} ,*

1. $\mathbf{X}_n + \mathbf{Y}_n \xrightarrow{d} \mathbf{X} + \mathbf{C}$
2. $\mathbf{Y}_n \mathbf{X}_n \xrightarrow{d} \mathbf{C} \mathbf{X}$
3. $\mathbf{Y}_n^{-1} \mathbf{X}_n \xrightarrow{d} \mathbf{C}^{-1} \mathbf{X}$ as long as \mathbf{C} is non-singular.

This theorem is at the core of hypothesis testing where estimated parameters are often asymptotically normal and an estimated parameter covariance, which converges in probability to the true covariance, is used to studentize the parameters.

2.3 Properties of Estimators

The first step in assessing the performance of an economic model is the estimation of the parameters. There are a number of desirable properties estimators may possess.

2.3.1 Bias and Consistency

A natural question to ask about an estimator is whether, on average, it will be equal to the population value of the parameter estimated. Any discrepancy between the expected value of an estimator and the population parameter is known as bias.

Definition 2.8 (Bias). The bias of an estimator, $\hat{\theta}$, is defined

$$B[\hat{\theta}] = E[\hat{\theta}] - \theta_0 \quad (2.15)$$

where θ_0 is used to denote the population (or “true”) value of the parameter.

When an estimator has a bias of 0 it is said to be unbiased. Unfortunately, many estimators are not unbiased. Consistency is a closely related concept that measures whether a parameter will be far from the population value in *large samples*.

Definition 2.9 (Consistency). An estimator $\hat{\theta}_n$ is said to be consistent if $\text{plim} \hat{\theta}_n = \theta_0$. The explicit dependence of the estimator on the sample size is used to clarify that these form a sequence, $\{\hat{\theta}_n\}_{n=1}^{\infty}$.

Consistency requires an estimator to exhibit two features as the sample size becomes large. First, any bias must be shrinking. Second, the distribution of $\hat{\theta}$ around θ_0 must be shrinking in such a way that virtually all of the probability mass is arbitrarily close to θ_0 . Behind consistency is a set of theorems known as *laws of large numbers*. Laws of large numbers provide conditions where an average will converge to its expectation. The simplest is the Kolmogorov Strong Law of Large numbers and is applicable to i.i.d. data.⁸

Theorem 2.6 (Kolmogorov Strong Law of Large Numbers). *Let $\{y_i\}$ by a sequence of i.i.d. random variables with $\mu \equiv E[y_i]$ and define $\bar{y}_n = n^{-1} \sum_{i=1}^n y_i$. Then*

$$\bar{y}_n \xrightarrow{a.s.} \mu \quad (2.16)$$

if and only if $E[|y_i|] < \infty$.

In the case of i.i.d. data the only requirement for consistency is that the expectation exists, and so a law of large numbers will apply to an average of i.i.d. data whenever its expectation exists. For example, Monte Carlo integration uses i.i.d. draws and so the Kolmogorov LLN is sufficient to ensure that Monte Carlo integrals converge to their expected values.

The variance of an estimator is the same as any other variance, $V[\hat{\theta}] = E[(\hat{\theta} - E[\hat{\theta}])^2]$ although it is worth noting that the variance is defined as the variation around its expectation, $E[\hat{\theta}]$, not the population value of the parameters, θ_0 . Mean square error measures this alternative form of variation around the population value of the parameter.

⁸A law of large numbers is strong if the convergence is almost sure. It is weak if convergence is in probability.

Definition 2.10 (Mean Square Error). The mean square error of an estimator $\hat{\theta}$, denoted $MSE(\hat{\theta})$, is defined

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta_0)^2]. \quad (2.17)$$

It can be equivalently expressed as the bias squared plus the variance, $MSE(\hat{\theta}) = B[\hat{\theta}]^2 + V[\hat{\theta}]$.

When the bias and variance of an estimator both converge to zero, then $\hat{\theta}_n \xrightarrow{m.s.} \theta_0$.

2.3.1.1 Bias and Consistency of the Method of Moment Estimators

The method of moments estimators of the mean and variance are defined as

$$\begin{aligned}\hat{\mu} &= n^{-1} \sum_{i=1}^n y_i \\ \hat{\sigma}^2 &= n^{-1} \sum_{i=1}^n (y_i - \hat{\mu})^2.\end{aligned}$$

When the data are i.i.d. with finite mean μ and variance σ^2 , the mean estimator is unbiased while the variance is biased by an amount that becomes small as the sample size increases. The mean is unbiased since

$$\begin{aligned}E[\hat{\mu}] &= E\left[n^{-1} \sum_{i=1}^n y_i\right] \\ &= n^{-1} \sum_{i=1}^n E[y_i] \\ &= n^{-1} \sum_{i=1}^n \mu \\ &= n^{-1} n \mu \\ &= \mu\end{aligned}$$

The variance estimator is biased since

$$\begin{aligned}E[\hat{\sigma}^2] &= E\left[n^{-1} \sum_{i=1}^n (y_i - \hat{\mu})^2\right] \\ &= E\left[n^{-1} \left(\sum_{i=1}^n y_i^2 - n\hat{\mu}^2 \right)\right] \\ &= n^{-1} \left(\sum_{i=1}^n E[y_i^2] - nE[\hat{\mu}^2] \right)\end{aligned}$$

$$\begin{aligned}
&= n^{-1} \left(\sum_{i=1}^n \mu^2 + \sigma^2 - n \left(\mu^2 + \frac{\sigma^2}{n} \right) \right) \\
&= n^{-1} (n\mu^2 + n\sigma^2 - n\mu^2 - \sigma^2) \\
&= n^{-1} \left(n\sigma^2 - n \frac{\sigma^2}{n} \right) \\
&= \frac{n-1}{n} \sigma^2
\end{aligned}$$

where the sample mean is equal to the population mean plus an error that is decreasing in n ,

$$\begin{aligned}
\hat{\mu}^2 &= \left(\mu + n^{-1} \sum_{i=1}^n \varepsilon_i \right)^2 \\
&= \mu^2 + 2\mu n^{-1} \sum_{i=1}^n \varepsilon_i + \left(n^{-1} \sum_{i=1}^n \varepsilon_i \right)^2
\end{aligned}$$

and so its square has the expected value

$$\begin{aligned}
E[\hat{\mu}^2] &= E \left[\mu^2 + 2\mu n^{-1} \sum_{i=1}^n \varepsilon_i + \left(n^{-1} \sum_{i=1}^n \varepsilon_i \right)^2 \right] \\
&= \mu^2 + 2\mu n^{-1} E \left[\sum_{i=1}^n \varepsilon_i \right] + n^{-2} E \left[\left(\sum_{i=1}^n \varepsilon_i \right)^2 \right] \\
&= \mu^2 + \frac{\sigma^2}{n}.
\end{aligned}$$

2.3.2 Asymptotic Normality

While unbiasedness and consistency are highly desirable properties of any estimator, alone these do not provide a method to perform inference. The primary tool in econometrics for inference is the central limit theorem (CLT). CLTs exist for a wide range of possible data characteristics that include i.i.d., heterogeneous and dependent cases. The Lindberg-Lévy CLT, which is applicable to i.i.d. data, is the simplest.

Theorem 2.7 (Lindberg-Lévy). *Let $\{y_i\}$ be a sequence of i.i.d. random scalars with $\mu \equiv E[Y_i]$ and $\sigma^2 \equiv V[Y_i] < \infty$. If $\sigma^2 > 0$, then*

$$\frac{\bar{y}_n - \mu}{\bar{\sigma}_n} = \sqrt{n} \frac{\bar{y}_n - \mu}{\sigma} \xrightarrow{d} N(0, 1) \quad (2.18)$$

where $\bar{y}_n = n^{-1} \sum_{i=1}^n y_i$ and $\bar{\sigma}_n = \sqrt{\frac{\sigma^2}{n}}$.

Lindberg-Lévy states that as long as i.i.d. data have 2 moments – a mean and variance – the sample mean will be asymptotically normal. It can further be seen to show that other moments of i.i.d. random variables, such as the variance, will be asymptotically normal as long as two times the power of the moment exists. In other words, an estimator of the r^{th} moment will be asymptotically normal as long as the $2r^{\text{th}}$ moment exists – at least in i.i.d. data. Figure 2.2 contains density plots of the sample average of n independent χ_1^2 random variables for $n = 5, 10, 50$ and 100 .⁹ The top panel contains the density of the unscaled estimates. The bottom panel contains the density plot of the correctly scaled terms, $\sqrt{n}(\hat{\mu} - 1)/\sqrt{2}$ where $\hat{\mu}$ is the sample average. The densities are collapsing in the top panel. This is evidence of consistency since the asymptotic distribution of $\hat{\mu}$ is collapsing on 1. The bottom panel demonstrates the operation of a CLT since the appropriately standardized means all have similar dispersion and are increasingly normal.

Central limit theorems exist for a wide variety of other data generating process including processes which are independent but not identically distributed (i.n.i.d) or processes which are dependent, such as time-series data. As the data become more heterogeneous, whether through dependence or by having different variance or distributions, more restrictions are needed on certain characteristics of the data to ensure that averages will be asymptotically normal. The Lindberg-Feller CLT allows for heteroskedasticity (different variances) and/or different marginal distributions.

Theorem 2.8 (Lindberg-Feller). *Let $\{y_i\}$ be a sequence of independent random scalars with $\mu_i \equiv E[y_i]$ and $0 < \sigma_i^2 \equiv V[y_i] < \infty$ where $y_i \sim F_i$, $i = 1, 2, \dots$. Then*

$$\sqrt{n} \frac{\bar{y}_n - \bar{\mu}_n}{\bar{\sigma}_n} \xrightarrow{d} N(0, 1) \quad (2.19)$$

and

$$\lim_{n \rightarrow \infty} \max_{1 \leq i \leq n} n^{-1} \frac{\sigma_i^2}{\bar{\sigma}_n^2} = 0 \quad (2.20)$$

if and only if, for every $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} \bar{\sigma}_n^2 n^{-1} \sum_{i=1}^n \int_{(z-\mu_n)^2 > \epsilon N \bar{\sigma}_n^2} (z - \mu_n)^2 dF_i(z) = 0 \quad (2.21)$$

where $\bar{\mu} = n^{-1} \sum_{i=1}^n \mu_i$ and $\bar{\sigma}^2 = n^{-1} \sum_{i=1}^n \sigma_i^2$.

The Lindberg-Feller CLT relaxes the requirement that the marginal distributions are identical in the Lindberg-Lévy CLT at the cost of a technical condition. The final condition, known as a Lindberg condition, essentially that no random variable is so heavy-tailed that it dominates the others when averaged. In practice, this can be a concern when the variables have a wide range of variances (σ_i^2). Many macroeconomic data series exhibit a large decrease in the variance of their shocks after 1984, a phenomenon is referred to as the *great moderation*. The statistical consequence of this decrease is that averages that use data both before and after 1984 not be well approximated by a CLT and caution is warranted when using asymptotic approximations. This phenomena is also present in equity returns where some periods – for example the technology “bubble” from 1997-2002 – have substantially higher volatility than periods before or after. These large persistent changes in the characteristics of the data have negative consequences on the quality of CLT approximations and large data samples are often needed.

⁹The mean and variance of a χ_v^2 are v and $2v$, respectively.

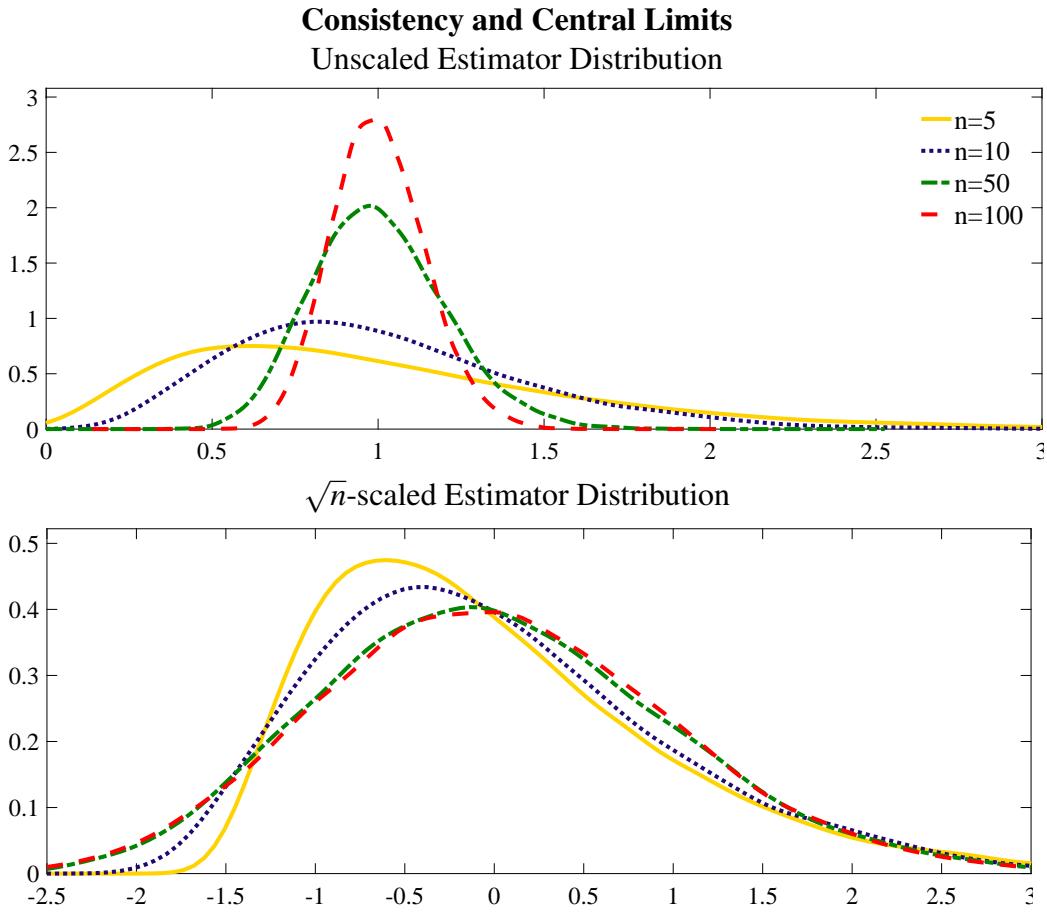


Figure 2.2: These two panels illustrate the difference between consistency and the correctly scaled estimators. The sample mean was computed 1,000 times using 5, 10, 50 and 100 i.i.d. χ^2 data points. The top panel contains a kernel density plot of the estimates of the mean. The density when $n = 100$ is much tighter than when $n = 5$ or $n = 10$ since the estimates are not scaled. The bottom panel plots $\sqrt{n}(\hat{\mu} - 1)/\sqrt{2}$, the standardized version for which a CLT applies. All scaled densities have similar dispersion although it is clear that the asymptotic approximation of the CLT is not particularly accurate when $n = 5$ or $n = 10$ due to the right skew in the χ^2_1 data.

2.3.2.1 What good is a CLT?

Central limit theorems are the basis of most inference in econometrics, although their formal justification is only asymptotic and hence only guaranteed to be valid for an arbitrarily large data set. Reconciling these two statements is an important step in the evolution of an econometrician.

Central limit theorems should be seen as approximations, and as an approximation, they can be accurate or arbitrarily poor. For example, when a series of random variables are i.i.d., thin-tailed and not skewed, the distribution of the sample mean computed using as few as 10 observations may be very well approximated using a central limit theorem. On the other hand, the approximation of a central limit theorem for the estimate of the autoregressive parameter, ρ , in

$$y_i = \rho y_{i-1} + \varepsilon_i \quad (2.22)$$

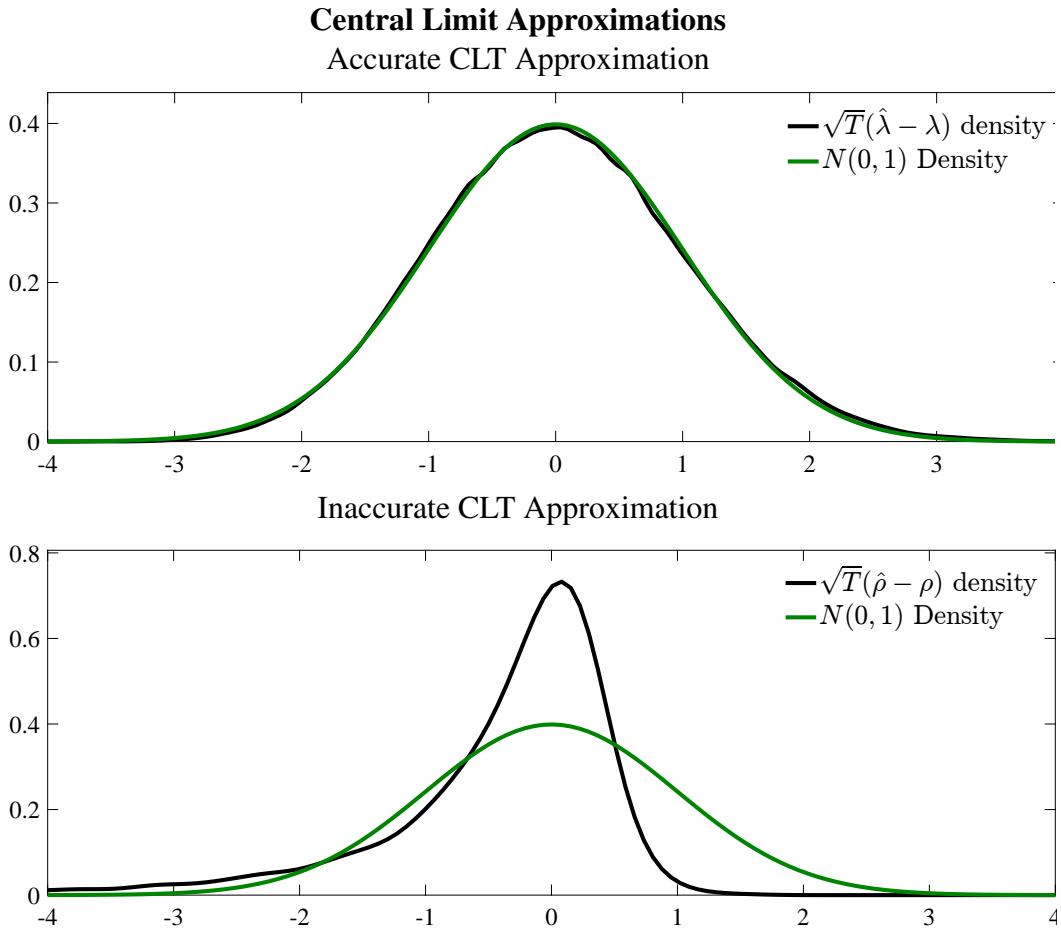


Figure 2.3: These two plots illustrate how a CLT can provide a good approximation, even in small samples (top panel), or a bad approximation even for moderately large samples (bottom panel). The top panel contains a kernel density plot of the standardized sample mean of $n = 10$ Poisson random variables ($\lambda = 5$) over 10,000 Monte Carlo simulations. Here the finite sample distribution and the asymptotic distribution overlay one another. The bottom panel contains the conditional ML estimates of ρ from the AR(1) $y_i = \rho y_{i-1} + \varepsilon_i$ where ε_i is i.i.d. standard normal using 100 data points and 10,000 replications. While $\hat{\rho}$ is asymptotically normal, the quality of the approximation when $n = 100$ is poor.

may be poor even for hundreds of data points when ρ is close to one (but smaller). Figure 2.3 contains kernel density plots of the sample means computed from a set of 10 i.i.d. draws from a Poisson distribution with $\lambda = 5$ in the top panel and the estimated autoregressive parameter from the autoregression in eq. (2.22) with $\rho = .995$ in the bottom. Each figure also contains the pdf of an appropriately scaled normal. The CLT for the sample means of the Poisson random variables is virtually indistinguishable from the actual distribution. On the other hand, the CLT approximation for $\hat{\rho}$ is very poor being based on 100 data points – $10\times$ more than in the i.i.d. uniform example. The difference arises because the data in the AR(1) example are not independent. With $\rho = 0.995$ data are highly dependent and more data is required for averages to be well behaved so that the CLT approximation is accurate.

There are no hard and fast rules as to when a CLT will be a good approximation. In general, the

more dependent and the more heterogeneous a series, the worse the approximation for a fixed number of observations. Simulations (Monte Carlo) are a useful tool to investigate the validity of a CLT since they allow the finite sample distribution to be tabulated and compared to the asymptotic distribution.

2.3.3 Efficiency

A final concept, efficiency, is useful for ranking consistent asymptotically normal (CAN) estimators that have the same rate of convergence.¹⁰

Definition 2.11 (Relative Efficiency). Let $\hat{\theta}_n$ and $\tilde{\theta}_n$ be two \sqrt{n} -consistent asymptotically normal estimators for θ_0 . If the asymptotic variance of $\hat{\theta}_n$, written $\text{avar}(\hat{\theta}_n)$ is less than the asymptotic variance of $\tilde{\theta}_n$, and so

$$\text{avar}(\hat{\theta}_n) < \text{avar}(\tilde{\theta}_n) \quad (2.23)$$

then $\hat{\theta}_n$ is said to be relatively efficient to $\tilde{\theta}_n$.¹¹

Note that when θ is a vector, $\text{avar}(\hat{\theta}_n)$ will be a covariance matrix. Inequality for matrices \mathbf{A} and \mathbf{B} is interpreted to mean that if $\mathbf{A} < \mathbf{B}$ then $\mathbf{B} - \mathbf{A}$ is positive semi-definite, and so *all* of the variances of the inefficient estimator must be (weakly) larger than those of the efficient estimator.

Definition 2.12 (Asymptotically Efficient Estimator). Let $\hat{\theta}_n$ and $\tilde{\theta}_n$ be two \sqrt{n} -consistent asymptotically normal estimators for θ_0 . If

$$\text{avar}(\hat{\theta}_n) < \text{avar}(\tilde{\theta}_n) \quad (2.24)$$

for any choice of $\tilde{\theta}_n$ then $\hat{\theta}_n$ is said to be the efficient estimator of θ .

One of the important features of efficiency comparisons is that they are only meaningful if both estimators are asymptotically normal, and hence consistent, at the same rate – \sqrt{n} in the usual case. It is trivial to produce an estimator that has a smaller variance but is inconsistent. For example, if an estimator for a scalar unknown is $\hat{\theta} = 7$ then it has no variance: it will always be 7. However, unless $\theta_0 = 7$ it will also be biased. Mean square error is a more appropriate method to compare estimators where one or more may be biased since it accounts for the total variation, not just the variance.¹²

2.4 Distribution Theory

Most distributional theory follows from a central limit theorem applied to the moment conditions or to the score of the log-The likelihood. While the moment conditions or scores are not usually the objects of interest – θ is – a simple expansion can be used to establish the asymptotic distribution of the estimated parameters.

¹⁰In any consistent estimator the asymptotic distribution of $\hat{\theta} - \theta_0$ is degenerate. In order to perform inference on an unknown quantity, the difference between the estimate and the population parameters must be scaled by a function of the number of data points. For most estimators this rate is \sqrt{n} , and so $\sqrt{n}(\hat{\theta} - \theta_0)$ will have an asymptotically normal distribution. In the general case, the scaled difference can be written as $n^\delta(\hat{\theta} - \theta_0)$ where n^δ is known as the rate.

¹¹The asymptotic variance of a \sqrt{n} -consistent estimator, written $\text{avar}(\hat{\theta}_n)$ is defined as $\lim_{n \rightarrow \infty} V[\sqrt{n}(\hat{\theta}_n - \theta_0)]$.

¹²Some consistent asymptotically normal estimators have an asymptotic bias and so $\sqrt{n}(\tilde{\theta}_n - \theta) \xrightarrow{d} N(\mathbf{B}, \Sigma)$. Asymptotic MSE defined as $E[n(\hat{\theta}_n - \theta_0)(\hat{\theta}_n - \theta_0)'] = \mathbf{B}\mathbf{B}' + \Sigma$ provides a method to compare estimators using their asymptotic properties.

2.4.1 Method of Moments

Distribution theory for the classical method of moments estimators is the most straightforward. Further, Maximum Likelihood can be considered a special case and so the method of moments is a natural starting point.¹³ The method of moments estimator is defined as

$$\begin{aligned}\hat{\mu} &= n^{-1} \sum_{i=1}^n x_i \\ \hat{\mu}_2 &= n^{-1} \sum_{i=1}^n (x_i - \hat{\mu})^2 \\ &\vdots \\ \hat{\mu}_k &= n^{-1} \sum_{i=1}^n (x_i - \hat{\mu})^k\end{aligned}$$

To understand the distribution theory for the method of moments estimator, begin by reformulating the estimator as the solution of a set of k equations evaluated using the population values of μ, μ_2, \dots

$$\begin{aligned}n^{-1} \sum_{i=1}^n x_i - \mu &= 0 \\ n^{-1} \sum_{i=1}^n (x_i - \mu)^2 - \mu_2 &= 0 \\ &\vdots \\ n^{-1} \sum_{i=1}^n (x_i - \mu)^k - \mu_k &= 0\end{aligned}$$

Define $g_{1i} = x_i - \mu$ and $g_{ji} = (x_i - \mu)^j - \mu_j$, $j = 2, \dots, k$, and the vector \mathbf{g}_i as

$$\mathbf{g}_i = \begin{bmatrix} g_{1i} \\ g_{2i} \\ \vdots \\ g_{ki} \end{bmatrix}. \quad (2.25)$$

Using this definition, the method of moments estimator can be seen as the solution to

$$\mathbf{g}_n(\hat{\theta}) = n^{-1} \sum_{i=1}^n \mathbf{g}_i(\hat{\theta}) = 0.$$

¹³While the class of method of moments estimators and maximum likelihood estimators contains a substantial overlap, method of moments estimators exist that cannot be replicated as a score condition of any likelihood since the likelihood is required to integrate to 1.

Consistency of the method of moments estimator relies on a *law of large numbers* holding for $n^{-1} \sum_{i=1}^n x_i$ and $n^{-1} \sum_{i=1}^n (x_i - \mu)^j$ for $j = 2, \dots, k$. If x_i is an i.i.d. sequence and as long as $E[|x_n - \mu|^j]$ exists, then $n^{-1} \sum_{i=1}^n (x_i - \mu)^j \xrightarrow{p} \mu_j$.¹⁴ An alternative, and more restrictive approach is to assume that $E[(x_n - \mu)^{2j}] = \mu_{2j}$ exists, and so

$$E\left[n^{-1} \sum_{i=1}^n (x_i - \mu)^j\right] = \mu_j \quad (2.26)$$

$$\begin{aligned} V\left[n^{-1} \sum_{i=1}^n (x_i - \mu)^j\right] &= n^{-1} \left(E\left[(x_i - \mu)^{2j}\right] - E\left[(x_i - \mu)^j\right]^2 \right) \\ &= n^{-1} (\mu_{2j} - \mu_j^2), \end{aligned} \quad (2.27)$$

and so $n^{-1} \sum_{i=1}^n (x_i - \mu)^j \xrightarrow{m.s.} \mu_j$ which implies consistency.

The asymptotic normality of parameters estimated using the method of moments follows from the asymptotic normality of

$$\sqrt{n} \left(n^{-1} \sum_{i=1}^n \mathbf{g}_i(\theta_0) \right) = n^{-1/2} \sum_{i=1}^n \mathbf{g}_i(\theta_0), \quad (2.28)$$

an assumption. This requires the elements of \mathbf{g}_n to be sufficiently well behaved so that averages are asymptotically normally distributed. For example, when x_i is i.i.d., the Lindeberg-Lévy CLT would require x_i to have $2k$ moments when estimating k parameters. When estimating the mean, 2 moments are required (i.e. the variance is finite). To estimate the mean and the variance using i.i.d. data, 4 moments are required for the estimators to follow a CLT. As long as the moment conditions are differentiable in the actual parameters of interest θ – for example, the mean and the variance – a *mean value expansion* can be used to establish the asymptotic normality of these parameters.¹⁵

$$\begin{aligned} n^{-1} \sum_{i=1}^n \mathbf{g}_i(\hat{\theta}) &= n^{-1} \sum_{i=1}^n \mathbf{g}_i(\theta_0) + n^{-1} \sum_{i=1}^n \frac{\partial \mathbf{g}_i(\theta)}{\partial \theta'} \Big|_{\theta=\bar{\theta}} (\hat{\theta} - \theta_0) \\ &= n^{-1} \sum_{i=1}^n \mathbf{g}_i(\theta_0) + \mathbf{G}_n(\bar{\theta})(\hat{\theta} - \theta_0) \end{aligned} \quad (2.30)$$

¹⁴Technically, $n^{-1} \sum_{i=1}^n (x_i - \mu)^j \xrightarrow{a.s.} \mu_j$ by the Kolmogorov law of large numbers, but since a.s. convergence implies convergence in probability, the original statement is also true.

¹⁵The mean value expansion is defined in the following theorem.

Theorem 2.9 (Mean Value Theorem). *Let $s : \mathbb{R}^k \rightarrow \mathbb{R}$ be defined on a convex set $\Theta \subset \mathbb{R}^k$. Further, let s be continuously differentiable on Θ with k by 1 gradient*

$$\nabla s(\hat{\theta}) \equiv \frac{\partial s(\theta)}{\partial \theta} \Big|_{\theta=\hat{\theta}}. \quad (2.29)$$

Then for any points θ and θ_0 there exists $\bar{\theta}$ lying on the segment between θ and θ_0 such that $s(\theta) = s(\theta_0) + \nabla s(\bar{\theta})(\theta - \theta_0)$.

where $\bar{\theta}$ is a vector that lies between $\hat{\theta}$ and θ_0 , element-by-element. Note that $n^{-1} \sum_{i=1}^n \mathbf{g}_i(\hat{\theta}) = \mathbf{0}$ by construction and so

$$\begin{aligned} n^{-1} \sum_{i=1}^n \mathbf{g}_i(\theta_0) + \mathbf{G}_n(\bar{\theta})(\hat{\theta} - \theta_0) &= 0 \\ \mathbf{G}_n(\bar{\theta})(\hat{\theta} - \theta_0) &= -n^{-1} \sum_{i=1}^n \mathbf{g}_i(\theta_0) \\ (\hat{\theta} - \theta_0) &= -\mathbf{G}_n(\bar{\theta})^{-1} n^{-1} \sum_{i=1}^n \mathbf{g}_i(\theta_0) \\ \sqrt{n}(\hat{\theta} - \theta_0) &= -\mathbf{G}_n(\bar{\theta})^{-1} \sqrt{n} n^{-1} \sum_{i=1}^n \mathbf{g}_i(\theta_0) \\ \sqrt{n}(\hat{\theta} - \theta_0) &= -\mathbf{G}_n(\bar{\theta})^{-1} \sqrt{n} \mathbf{g}_n(\theta_0) \end{aligned}$$

where $\mathbf{g}_n(\theta_0) = n^{-1} \sum_{i=1}^n \mathbf{g}_i(\theta_0)$ is the average of the moment conditions. Thus the normalized difference between the estimated and the population values of the parameters, $\sqrt{n}(\hat{\theta} - \theta_0)$ is equal to a scaled $(-\mathbf{G}_n(\bar{\theta})^{-1})$ random variable ($\sqrt{n} \mathbf{g}_n(\theta_0)$) that has an asymptotic normal distribution. By assumption $\sqrt{n} \mathbf{g}_n(\theta_0) \xrightarrow{d} N(0, \Sigma)$ and so

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{G}^{-1} \Sigma (\mathbf{G}')^{-1}\right) \quad (2.31)$$

where $\mathbf{G}_n(\bar{\theta})$ has been replaced with its limit as $n \rightarrow \infty$, \mathbf{G} .

$$\begin{aligned} \mathbf{G} &= \text{plim}_{n \rightarrow \infty} \frac{\partial \mathbf{g}_n(\theta)}{\partial \theta'} \Big|_{\theta=\theta_0} \\ &= \text{plim}_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \frac{\partial \mathbf{g}_i(\theta)}{\partial \theta'} \Big|_{\theta=\theta_0} \end{aligned} \quad (2.32)$$

Since $\hat{\theta}$ is a consistent estimator, $\hat{\theta} \xrightarrow{p} \theta_0$ and so $\bar{\theta} \xrightarrow{p} \theta_0$ since it is between $\hat{\theta}$ and θ_0 . This form of asymptotic covariance is known as a “sandwich” covariance estimator.

2.4.1.1 Inference on the Mean and Variance

To estimate the mean and variance by the method of moments, two moment conditions are needed,

$$\begin{aligned} n^{-1} \sum_{i=1}^n x_i &= \hat{\mu} \\ n^{-1} \sum_{i=1}^n (x_i - \hat{\mu})^2 &= \hat{\sigma}^2 \end{aligned}$$

To derive the asymptotic distribution, begin by forming \mathbf{g}_i ,

$$\mathbf{g}_i = \begin{bmatrix} x_i - \mu \\ (x_i - \mu)^2 - \sigma^2 \end{bmatrix}$$

Note that \mathbf{g}_i is mean 0 and a function of a single x_i so that \mathbf{g}_i is also i.i.d.. The covariance of \mathbf{g}_i is given by

$$\begin{aligned} \Sigma &= E[\mathbf{g}_i \mathbf{g}_i'] = E\left[\begin{bmatrix} x_i - \mu \\ (x_i - \mu)^2 - \sigma^2 \end{bmatrix} \begin{bmatrix} x_i - \mu & (x_i - \mu)^2 - \sigma^2 \end{bmatrix}\right] \quad (2.33) \\ &= E\left[\begin{bmatrix} (x_i - \mu)^2 & (x_i - \mu)((x_i - \mu)^2 - \sigma^2) \\ (x_i - \mu)((x_i - \mu)^2 - \sigma^2) & ((x_i - \mu)^2 - \sigma^2)^2 \end{bmatrix}\right] \\ &= E\left[\begin{bmatrix} (x_i - \mu)^2 & (x_i - \mu)^3 - \sigma^2(x_i - \mu) \\ (x_i - \mu)^3 - \sigma^2(x_i - \mu) & (x_i - \mu)^4 - 2\sigma^2(x_i - \mu)^2 + \sigma^4 \end{bmatrix}\right] \\ &= \begin{bmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{bmatrix} \end{aligned}$$

and the Jacobian is

$$\begin{aligned} \mathbf{G} &= \text{plim}_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \frac{\partial \mathbf{g}_i(\theta)}{\partial \theta'} \Big|_{\theta=\theta_0} \\ &= \text{plim}_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \begin{bmatrix} -1 & 0 \\ -2(x_i - \mu) & -1 \end{bmatrix}. \end{aligned}$$

Since $\text{plim}_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n (x_i - \mu) = \text{plim}_{n \rightarrow \infty} \bar{x}_n - \mu = 0$,

$$\mathbf{G} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Thus, the asymptotic distribution of the method of moments estimator of $\theta = (\mu, \sigma^2)'$ is

$$\sqrt{n} \left(\begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{bmatrix} - \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} \right) \xrightarrow{d} N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{bmatrix} \right)$$

since $\mathbf{G} = -\mathbf{I}_2$ and so $\mathbf{G}^{-1} \Sigma (\mathbf{G}^{-1})' = -\mathbf{I}_2 \Sigma (-\mathbf{I}_2) = \Sigma$.

2.4.2 Maximum Likelihood

The steps to deriving the asymptotic distribution of a ML estimator are similar to those for a method of moments estimator where the score of the likelihood takes the place of the moment conditions. The maximum likelihood estimator is defined as the maximum of the log-likelihood of the data with respect to the parameters,

$$\hat{\theta} = \arg \max_{\theta} l(\theta; \mathbf{y}). \quad (2.34)$$

When the data are i.i.d., the log-likelihood can be factored into n log-likelihoods, one for each observation¹⁶,

$$l(\theta; \mathbf{y}) = \sum_{i=1}^n l_i(\theta; y_i). \quad (2.35)$$

It is useful to work with the average log-likelihood directly, and so define

$$\bar{l}_n(\theta; \mathbf{y}) = n^{-1} \sum_{i=1}^n l_i(\theta; y_i). \quad (2.36)$$

The intuition behind the asymptotic distribution follows from the use of the average. Under some regularity conditions, $\bar{l}_n(\theta; \mathbf{y})$ converges uniformly in θ to $E[l(\theta; y_i)]$. However, since the average log-likelihood is becoming a good approximation for the expectation of the log-likelihood, the value of θ that maximizes the log-likelihood of the data and its expectation will be very close for n sufficiently large. As a result, whenever the log-likelihood is differentiable and the range of y_i does not depend on any of the parameters in θ ,

$$E \left[\frac{\partial \bar{l}_n(\theta; y_i)}{\partial \theta} \Big|_{\theta=\theta_0} \right] = \mathbf{0} \quad (2.37)$$

where θ_0 are the parameters of the data generating process. This follows since

$$\begin{aligned} \int_{\mathcal{S}_{\mathbf{y}}} \frac{\partial \bar{l}_n(\theta_0; \mathbf{y})}{\partial \theta} \Big|_{\theta=\theta_0} f(\mathbf{y}; \theta_0) dy &= \int_{\mathcal{S}_{\mathbf{y}}} \frac{\frac{\partial f(\mathbf{y}; \theta_0)}{\partial \theta}}{f(\mathbf{y}; \theta_0)} \Big|_{\theta=\theta_0} f(\mathbf{y}; \theta_0) dy \\ &= \int_{\mathcal{S}_{\mathbf{y}}} \frac{\partial f(\mathbf{y}; \theta_0)}{\partial \theta} \Big|_{\theta=\theta_0} dy \\ &= \frac{\partial}{\partial \theta} \int_{\mathcal{S}_{\mathbf{y}}} f(\mathbf{y}; \theta) \Big|_{\theta=\theta_0} dy \\ &= \frac{\partial}{\partial \theta} 1 \\ &= 0 \end{aligned} \quad (2.38)$$

where $\mathcal{S}_{\mathbf{y}}$ denotes the support of \mathbf{y} . The scores of the average log-likelihood are

¹⁶Even when the data are not i.i.d., the log-likelihood can be factored into n log-likelihoods using conditional distributions for y_2, \dots, y_i and the marginal distribution of y_1 ,

$$l(\theta; \mathbf{y}) = \sum_{n=2}^N l_i(\theta; y_i | y_{i-1}, \dots, y_1) + l_1(\theta; y_1).$$

$$\frac{\partial \bar{l}_n(\theta; y_i)}{\partial \theta} = n^{-1} \sum_{i=1}^n \frac{\partial l_i(\theta; y_i)}{\partial \theta} \quad (2.39)$$

and when y_i is i.i.d. the scores will be i.i.d., and so the average scores will follow a law of large numbers for θ close to θ_0 . Thus

$$n^{-1} \sum_{i=1}^n \frac{\partial l_i(\theta; y_i)}{\partial \theta} \xrightarrow{a.s.} E\left[\frac{\partial l(\theta; Y_i)}{\partial \theta}\right] \quad (2.40)$$

As a result, the population value of θ , θ_0 , will also asymptotically solve the first order condition. The average scores are also the basis of the asymptotic normality of maximum likelihood estimators. Under some further regularity conditions, the average scores will follow a central limit theorem, and so

$$\sqrt{n} \nabla_{\theta} \bar{l}(\theta_0) \equiv \sqrt{n} \left(n^{-1} \sum_{i=1}^n \frac{\partial l(\theta; y_i)}{\partial \theta} \right) \Big|_{\theta=\theta_0} \xrightarrow{d} N(\mathbf{0}, \mathcal{J}). \quad (2.41)$$

Taking a mean value expansion around θ_0 ,

$$\begin{aligned} \sqrt{n} \nabla_{\theta} \bar{l}(\hat{\theta}) &= \sqrt{n} \nabla_{\theta} \bar{l}(\theta_0) + \sqrt{n} \nabla_{\theta \theta'} \bar{l}(\bar{\theta})(\hat{\theta} - \theta_0) \\ \mathbf{0} &= \sqrt{n} \nabla_{\theta} \bar{l}(\theta_0) + \sqrt{n} \nabla_{\theta \theta'} \bar{l}(\bar{\theta})(\hat{\theta} - \theta_0) \\ -\sqrt{n} \nabla_{\theta \theta'} \bar{l}(\bar{\theta})(\hat{\theta} - \theta_0) &= \sqrt{n} \nabla_{\theta} \bar{l}(\theta_0) \\ \sqrt{n}(\hat{\theta} - \theta_0) &= [-\nabla_{\theta \theta'} \bar{l}(\bar{\theta})]^{-1} \sqrt{n} \nabla_{\theta} l(\theta_0) \end{aligned}$$

where

$$\nabla_{\theta \theta'} \bar{l}(\bar{\theta}) \equiv n^{-1} \sum_{i=1}^n \frac{\partial^2 l(\theta; y_i)}{\partial \theta \partial \theta'} \Big|_{\theta=\bar{\theta}} \quad (2.42)$$

and where $\bar{\theta}$ is a vector whose elements lie between $\hat{\theta}$ and θ_0 . Since $\hat{\theta}$ is a consistent estimator of θ_0 , $\bar{\theta} \xrightarrow{P} \theta_0$ and so functions of $\bar{\theta}$ will converge to their value at θ_0 , and the asymptotic distribution of the maximum likelihood estimator is

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}^{-1} \mathcal{J} \mathcal{I}^{-1}) \quad (2.43)$$

where

$$\mathcal{I} = -E\left[\frac{\partial^2 l(\theta; y_i)}{\partial \theta \partial \theta'} \Big|_{\theta=\theta_0}\right] \quad (2.44)$$

and

$$\mathcal{J} = E\left[\frac{\partial l(\theta; y_i)}{\partial \theta} \frac{\partial l(\theta; y_i)}{\partial \theta'} \Big|_{\theta=\theta_0}\right] \quad (2.45)$$

The asymptotic covariance matrix can be further simplified using the information matrix equality which states that $\mathcal{I} - \mathcal{J} \xrightarrow{P} \mathbf{0}$ and so

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}^{-1}) \quad (2.46)$$

or equivalently

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(\mathbf{0}, \mathcal{J}^{-1}). \quad (2.47)$$

The information matrix equality follows from taking the derivative of the expected score,

$$\begin{aligned} \frac{\partial^2 l(\theta_0; \mathbf{y})}{\partial \theta \partial \theta'} &= \frac{1}{f(\mathbf{y}; \theta)} \frac{\partial^2 f(\mathbf{y}; \theta_0)}{\partial \theta \partial \theta'} - \frac{1}{f(\mathbf{y}; \theta)^2} \frac{\partial f(\mathbf{y}; \theta_0)}{\partial \theta} \frac{\partial f(\mathbf{y}; \theta_0)}{\partial \theta'} \\ \frac{\partial^2 l(\theta_0; \mathbf{y})}{\partial \theta \partial \theta'} + \frac{\partial l(\theta_0; \mathbf{y})}{\partial \theta} \frac{\partial l(\theta_0; \mathbf{y})}{\partial \theta'} &= \frac{1}{f(\mathbf{y}; \theta)} \frac{\partial^2 f(\mathbf{y}; \theta_0)}{\partial \theta \partial \theta'} \end{aligned} \quad (2.48)$$

and so, when the model is correctly specified,

$$\begin{aligned} \mathbb{E}\left[\frac{\partial^2 l(\theta_0; \mathbf{y})}{\partial \theta \partial \theta'} + \frac{\partial l(\theta_0; \mathbf{y})}{\partial \theta} \frac{\partial l(\theta_0; \mathbf{y})}{\partial \theta'}\right] &= \int_{S_y} \frac{1}{f(\mathbf{y}; \theta)} \frac{\partial^2 f(\mathbf{y}; \theta_0)}{\partial \theta \partial \theta'} f(\mathbf{y}; \theta) d\mathbf{y} \\ &= \int_{S_y} \frac{\partial^2 f(\mathbf{y}; \theta_0)}{\partial \theta \partial \theta'} d\mathbf{y} \\ &= \frac{\partial^2}{\partial \theta \partial \theta'} \int_{S_y} f(\mathbf{y}; \theta_0) d\mathbf{y} \\ &= \frac{\partial^2}{\partial \theta \partial \theta'} 1 \\ &= 0. \end{aligned}$$

and

$$\mathbb{E}\left[\frac{\partial^2 l(\theta_0; \mathbf{y})}{\partial \theta \partial \theta'}\right] = -\mathbb{E}\left[\frac{\partial l(\theta_0; \mathbf{y})}{\partial \theta} \frac{\partial l(\theta_0; \mathbf{y})}{\partial \theta'}\right].$$

A related concept, and one which applies to ML estimators when the information matrix equality holds – at least asymptotically – is the Cramér-Rao lower bound.

Theorem 2.10 (Cramér-Rao Inequality). *Let $f(\mathbf{y}; \theta)$ be the joint density of \mathbf{y} where θ is a k dimensional parameter vector. Let $\hat{\theta}$ be a consistent estimator of θ with finite covariance. Under some regularity condition on $f(\cdot)$*

$$\text{avar}(\hat{\theta}) \geq \mathcal{I}^{-1}(\theta) \quad (2.49)$$

where

$$\mathcal{I}(\theta) = -\mathbb{E}\left[\frac{\partial^2 \ln f(Y_i; \theta)}{\partial \theta \partial \theta'}\Big|_{\theta=\theta_0}\right]. \quad (2.50)$$

The important implication of the Cramér-Rao theorem is that maximum likelihood estimators, which are generally consistent, are asymptotically efficient.¹⁷ This guarantee makes a strong case for using the maximum likelihood when available.

2.4.2.1 Inference in a Poisson MLE

Recall that the log-likelihood in a Poisson MLE is

$$l(\lambda; \mathbf{y}) = -n\lambda + \ln(\lambda) \sum_{i=1}^n y_i - \sum_{i=1}^{y_i} \ln(i)$$

and that the first order condition is

$$\frac{\partial l(\lambda; \mathbf{y})}{\partial \lambda} = -n + \lambda^{-1} \sum_{i=1}^n y_i.$$

The MLE was previously shown to be $\hat{\lambda} = n^{-1} \sum_{i=1}^n y_i$. To compute the variance, take the expectation of the negative of the second derivative,

$$\frac{\partial^2 l(\lambda; y_i)}{\partial \lambda^2} = -\lambda^{-2}$$

and so

$$\begin{aligned} \mathcal{I} &= -E\left[\frac{\partial^2 l(\lambda; y_i)}{\partial \lambda^2}\right] = -E[-\lambda^{-2}] \\ &= [\lambda^{-2} E[y_i]] \\ &= [\lambda^{-2} \lambda] \\ &= \left[\frac{\lambda}{\lambda^2}\right] \\ &= \lambda^{-1} \end{aligned}$$

and so $\sqrt{n}(\hat{\lambda} - \lambda_0) \xrightarrow{d} N(0, \lambda)$ since $\mathcal{I}^{-1} = \lambda$.

Alternatively the covariance of the scores could be used to compute the parameter covariance,

$$\begin{aligned} \mathcal{J} &= V\left[\left(-1 + \frac{y_i}{\lambda}\right)^2\right] \\ &= \frac{1}{\lambda^2} V[y_i] \\ &= \lambda^{-1}. \end{aligned}$$

$\mathcal{I} = \mathcal{J}$ and so the IME holds when the data are Poisson distributed. If the data were not Poisson distributed, then it would not normally be the case that $E[y_i] = V[y_i] = \lambda$, and so \mathcal{I} and \mathcal{J} would not (generally) be equal.

¹⁷The Cramér-Rao bound also applied in finite samples when $\hat{\theta}$ is unbiased. While most maximum likelihood estimators are biased in finite samples, there are important cases where estimators are unbiased for any sample size and so the Cramér-Rao theorem will apply in finite samples. Linear regression is an important case where the Cramér-Rao theorem applies in finite samples (under some strong assumptions).

2.4.2.2 Inference in the Normal (Gaussian) MLE

Recall that the MLE estimators of the mean and variance are

$$\hat{\mu} = n^{-1} \sum_{i=1}^n y_i$$

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (y_i - \hat{\mu})^2$$

and that the log-likelihood is

$$l(\theta; \mathbf{y}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2} \sum_{i=1}^n \frac{(y_i - \mu)^2}{\sigma^2}.$$

Taking the derivative with respect to the parameter vector, $\theta = (\mu, \sigma^2)',$

$$\frac{\partial l(\theta; \mathbf{y})}{\partial \mu} = \sum_{i=1}^n \frac{(y_i - \mu)}{\sigma^2}$$

$$\frac{\partial l(\theta; \mathbf{y})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2} \sum_{i=1}^n \frac{(y_i - \mu)^2}{\sigma^4}.$$

The second derivatives are

$$\frac{\partial^2 l(\theta; \mathbf{y})}{\partial \mu \partial \mu} = -\sum_{i=1}^n \frac{1}{\sigma^2}$$

$$\frac{\partial^2 l(\theta; \mathbf{y})}{\partial \mu \partial \sigma^2} = -\sum_{i=1}^n \frac{(y_i - \mu)}{\sigma^4}$$

$$\frac{\partial^2 l(\theta; \mathbf{y})}{\partial \sigma^2 \partial \sigma^2} = \frac{n}{2\sigma^4} - \frac{2}{2} \sum_{i=1}^n \frac{(y_i - \mu)^2}{\sigma^6}.$$

The first does not depend on data and so no expectation is needed. The other two have expectations,

$$\begin{aligned} \mathbb{E} \left[\frac{\partial^2 l(\theta; y_i)}{\partial \mu \partial \sigma^2} \right] &= \mathbb{E} \left[-\frac{(y_i - \mu)}{\sigma^4} \right] \\ &= -\frac{(\mathbb{E}[y_i] - \mu)}{\sigma^4} \\ &= -\frac{\mu - \mu}{\sigma^4} \\ &= 0 \end{aligned}$$

and

$$\begin{aligned}
E \left[\frac{\partial^2 l(\theta; y_i)}{\partial \sigma^2 \partial \sigma^2} \right] &= E \left[\frac{1}{2\sigma^4} - \frac{2}{2} \frac{(y_i - \mu)^2}{\sigma^6} \right] \\
&= \frac{1}{2\sigma^4} - \frac{E[(y_i - \mu)^2]}{\sigma^6} \\
&= \frac{1}{2\sigma^4} - \frac{\sigma^2}{\sigma^6} \\
&= \frac{1}{2\sigma^4} - \frac{1}{\sigma^4} \\
&= -\frac{1}{2\sigma^4}
\end{aligned}$$

Putting these together, the expected Hessian can be formed,

$$E \left[\frac{\partial^2 l(\theta; y_i)}{\partial \theta \partial \theta'} \right] = \begin{bmatrix} -\frac{1}{\sigma^2} & 0 \\ 0 & -\frac{1}{2\sigma^4} \end{bmatrix}$$

and so the asymptotic covariance is

$$\begin{aligned}
\mathcal{I}^{-1} &= -E \left[\frac{\partial^2 l(\theta; y_i)}{\partial \theta \partial \theta'} \right]^{-1} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}^{-1} \\
&= \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix}
\end{aligned}$$

The asymptotic distribution is then

$$\sqrt{n} \left(\begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{bmatrix} - \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} \right) \xrightarrow{d} N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix} \right)$$

Note that this is different from the asymptotic variance for the method of moments estimator of the mean and the variance. This is because the data have been assumed to come from a normal distribution and so the MLE is correctly specified. As a result $\mu_3 = 0$ (the normal is symmetric) and the IME holds. In general the IME does not hold and so the asymptotic covariance may take a different form which depends on the moments of the data as in eq. (2.33).

2.4.3 Quasi Maximum Likelihood

While maximum likelihood is an appealing estimation approach, it has one important drawback: knowledge of $f(\mathbf{y}; \theta)$. In practice the density assumed in maximum likelihood estimation, $f(\mathbf{y}; \theta)$, is misspecified for the actual density of \mathbf{y} , $g(\mathbf{y})$. This case has been widely studied and estimators where the distribution is misspecified are known as quasi-maximum likelihood (QML) estimators. QML estimators generally lose all of the features that make maximum likelihood estimators so appealing: they are generally inconsistent for the parameters of interest, the information matrix equality does not hold and they do not achieve the Cramér-Rao lower bound.

First, consider the expected score from a QML estimator,

$$\begin{aligned}
 E_g \left[\frac{\partial l(\theta_0; \mathbf{y})}{\partial \theta} \right] &= \int_{S_y} \frac{\partial l(\theta_0; \mathbf{y})}{\partial \theta} g(\mathbf{y}) d\mathbf{y} \\
 &= \int_{S_y} \frac{\partial l(\theta_0; \mathbf{y})}{\partial \theta} \frac{f(\mathbf{y}; \theta_0)}{f(\mathbf{y}; \theta_0)} g(\mathbf{y}) d\mathbf{y} \\
 &= \int_{S_y} \frac{\partial l(\theta_0; \mathbf{y})}{\partial \theta} \frac{g(\mathbf{y})}{f(\mathbf{y}; \theta_0)} f(\mathbf{y}; \theta_0) d\mathbf{y} \\
 &= \int_{S_y} h(\mathbf{y}) \frac{\partial l(\theta_0; \mathbf{y})}{\partial \theta} f(\mathbf{y}; \theta_0) d\mathbf{y}
 \end{aligned} \tag{2.51}$$

which shows that the QML estimator can be seen as a weighted average with respect to the density assumed. However these weights depend on the data, and so it will no longer be the case that the expectation of the score at θ_0 will necessarily be 0. Instead QML estimators generally converge to another value of θ , θ^* , that depends on both $f(\cdot)$ and $g(\cdot)$ and is known as the pseudo-true value of θ .

The other important consideration when using QML to estimate parameters is that the Information Matrix Equality (IME) no longer holds, and so “sandwich” covariance estimators must be used and likelihood ratio statistics will not have standard χ^2 distributions. An alternative interpretation of a QML estimator is that of a method of moments estimator where the scores of $l(\theta; \mathbf{y})$ are used to choose the moments. With this interpretation, the distribution theory of the method of moments estimator will apply as long as the scores, evaluated at the pseudo-true parameters, follow a CLT.

2.4.3.1 The Effect of the Data Distribution on Estimated Parameters

Figure 2.4 contains three distributions (left column) and the asymptotic covariance of the mean and the variance estimators, illustrated through joint confidence ellipses containing 80, 95 and 99% probability the true value is within their bounds (right column).¹⁸ The ellipses were all derived from the asymptotic covariance of $\hat{\mu}$ and $\hat{\sigma}^2$ where the data are i.i.d. and distributed according to a *mixture of normals* distribution where

$$y_i = \begin{cases} \mu_1 + \sigma_1 z_i & \text{with probability } p \\ \mu_2 + \sigma_2 z_i & \text{with probability } 1 - p \end{cases}$$

where z is a standard normal. A mixture of normals is constructed from mixing draws from a finite set of normals with possibly different means and/or variances and can take a wide variety of shapes. All of the variables were constructed so that $E[y_i] = 0$ and $V[y_i] = 1$. This requires

$$p\mu_1 + (1 - p)\mu_2 = 0$$

and

$$p(\mu_1^2 + \sigma_1^2) + (1 - p)(\mu_2^2 + \sigma_2^2) = 1.$$

¹⁸The ellipses are centered at (0,0) since the population value of the parameters has been subtracted. Also note that even though the confidence ellipse for $\hat{\sigma}^2$ extended into the negative space, these must be divided by \sqrt{n} and re-centered at the estimated value when used.

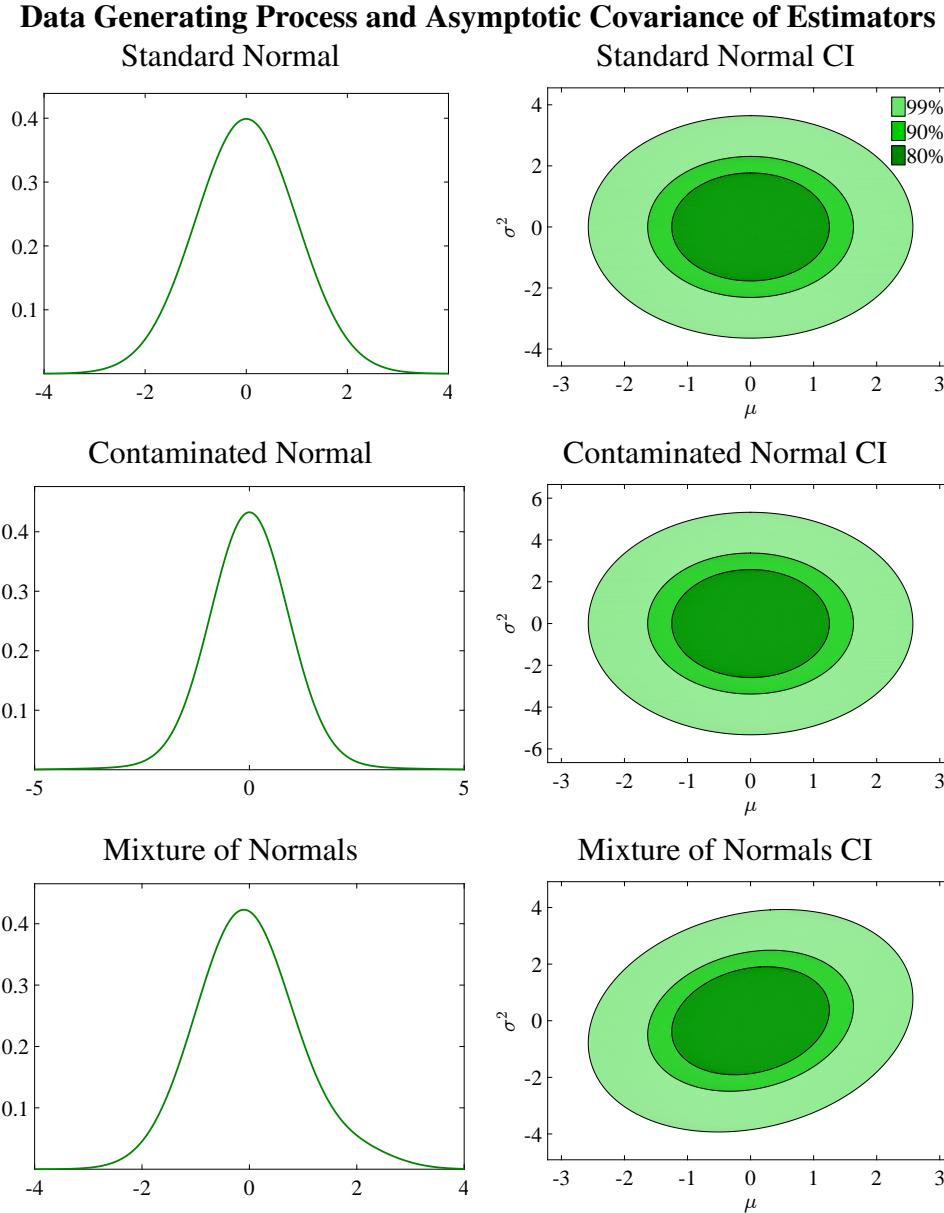


Figure 2.4: The six subplots illustrate how the data generating process, not the assumed model, determine the asymptotic covariance of parameter estimates. In each row of panels, the left shows the distribution of the data from a mixture of normals, $y_i = \mu_1 + \sigma_1 z_i$ with probability p and $y_i = \mu_2 + \sigma_2 z_i$ with probability $1 - p$. The right shows the asymptotic distribution of $\hat{\mu}$ and $\hat{\sigma}^2$. The parameters were chosen so that $E[y_i] = 0$ and $V[y_i] = 1$. Different parameter configurations produce a standard normal (top), a heavy tailed distribution known as a contaminated normal (middle) and a skewed distribution (bottom).

	p	μ_1	σ_1^2	μ_2	σ_2^2
Standard Normal	1	0	1	0	1
Contaminated Normal	.95	0	.8	0	4.8
Right Skewed Mixture	.05	2	.5	-.1	.8

Table 2.1: Parameter values used in the mixtures of normals illustrated in figure 2.4.

The values used to produce the figures are listed in table 2.1. The first set is simply a standard normal since $p = 1$. The second is known as a contaminated normal and is composed of a frequently occurring (95% of the time) mean-zero normal with variance slightly smaller than 1 (.8), contaminated by a rare but high variance (4.8) mean-zero normal. This produces heavy tails but does not result in a skewed distribution. The final example uses different means and variance to produce a right (positively) skewed distribution.

The confidence ellipses illustrated in figure 2.4 are all derived from estimators produced assuming that the data are normal, but using the “sandwich” version of the covariance, $\mathcal{I}^{-1}\mathcal{J}\mathcal{I}^{-1}$. The top panel illustrates the correctly specified maximum likelihood estimator. Here the confidence ellipse is symmetric about its center. This illustrates that the parameters are uncorrelated – and hence independent, since they are asymptotically normal – and that they have different variances. The middle panel has a similar shape but is elongated on the variance axis (x). This illustrates that the asymptotic variance of $\hat{\sigma}^2$ is affected by the heavy tails of the data (large 4th moment) of the contaminated normal. The final confidence ellipse is rotated which reflects that the mean and variance estimators are no longer asymptotically independent. These final two cases are examples of QML; the estimator is derived assuming a normal distribution but the data are not. In these examples, the estimators are still consistent but have different covariances.¹⁹

2.4.4 The Delta Method

Some theories make predictions about *functions* of parameters rather than on the parameters directly. One common example in finance is the Sharpe ratio, S , defined

$$S = \frac{\mathbb{E}[r - r_f]}{\sqrt{\text{V}[r - r_f]}} \quad (2.52)$$

where r is the return on a risky asset and r_f is the risk-free rate – and so $r - r_f$ is the excess return on the risky asset. While the quantities in both the numerator and the denominator are standard statistics, the mean and the standard deviation, the ratio is not.

The delta method can be used to compute the covariance of functions of asymptotically normal parameter estimates.

Definition 2.13 (Delta method). Let $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{G}^{-1}\Sigma(\mathbf{G}')^{-1}\right)$ where Σ is a positive definite covariance matrix. Further, suppose that $\mathbf{d}(\theta)$ is a m by 1 continuously differentiable vector function

¹⁹While these examples are consistent, it is not generally the case that the parameters estimated using a misspecified likelihood (QML) are consistent for the quantities of interest.

of θ from $\mathbb{R}^k \rightarrow \mathbb{R}^m$. Then,

$$\sqrt{n}(\mathbf{d}(\hat{\theta}) - \mathbf{d}(\theta_0)) \xrightarrow{d} N\left(0, \mathbf{D}(\theta_0) \left[\mathbf{G}^{-1} \Sigma (\mathbf{G}')^{-1} \right] \mathbf{D}(\theta_0)' \right)$$

where

$$\mathbf{D}(\theta_0) = \frac{\partial \mathbf{d}(\theta)}{\partial \theta'} \Big|_{\theta=\theta_0}. \quad (2.53)$$

2.4.4.1 Variance of the Sharpe Ratio

The Sharpe ratio is estimated by “plugging in” the usual estimators of the mean and the variance,

$$\hat{S} = \frac{\hat{\mu}}{\sqrt{\hat{\sigma}^2}}.$$

In this case $\mathbf{d}(\theta_0)$ is a scalar function of two parameters, and so

$$\mathbf{d}(\theta_0) = \frac{\mu}{\sqrt{\sigma^2}}$$

and

$$\mathbf{D}(\theta_0) = \begin{bmatrix} \frac{1}{\sigma} & \frac{-\mu}{2\sigma^3} \end{bmatrix}$$

Recall that the asymptotic distribution of the estimated mean and variance is

$$\sqrt{n} \left(\begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{bmatrix} - \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} \right) \xrightarrow{d} N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{bmatrix} \right).$$

The asymptotic distribution of the Sharpe ratio can be constructed by combining the asymptotic distribution of $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)'$ with the $\mathbf{D}(\theta_0)$, and so

$$\sqrt{n}(\hat{S} - S) \xrightarrow{d} N \left(0, \begin{bmatrix} \frac{1}{\sigma} & \frac{-\mu}{2\sigma^3} \end{bmatrix} \begin{bmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma} & \frac{-\mu}{2\sigma^3} \end{bmatrix}' \right)$$

which can be simplified to

$$\sqrt{n}(\hat{S} - S) \xrightarrow{d} N \left(0, 1 - \frac{\mu\mu_3}{\sigma^4} + \frac{\mu^2(\mu_4 - \sigma^4)}{4\sigma^6} \right).$$

The asymptotic variance can be rearranged to provide some insight into the sources of uncertainty,

$$\sqrt{n}(\hat{S} - S) \xrightarrow{d} N \left(0, 1 - S \times sk + \frac{1}{4}S^2(\kappa - 1) \right),$$

where sk is the skewness and κ is the kurtosis. This shows that the variance of the Sharpe ratio will be higher when the data is negatively skewed or when the data has a large kurtosis (heavy tails), both empirical regularities of asset pricing data. If asset returns were normally distributed, and so $sk = 0$ and $\kappa = 3$, the expression of the asymptotic variance simplifies to

$$\text{V} [\sqrt{n} (\hat{S} - S)] = 1 + \frac{S^2}{2}, \quad (2.54)$$

which is expression commonly used as the variance of the Sharpe ratio. As this example illustrates the expression in eq. (2.54) is *only* correct if the skewness is 0 and returns have a kurtosis of 3 – something that would only be expected if returns are normal.

2.4.5 Estimating Covariances

The presentation of the asymptotic theory in this chapter does not provide a method to implement hypothesis tests since all of the distributions depend on the covariance of the scores and the expected second derivative or Jacobian in the method of moments. Feasible testing requires estimates of these. The usual method to estimate the covariance uses “plug-in” estimators. Recall that in the notation of the method of moments,

$$\Sigma \equiv \text{avar} \left(n^{-\frac{1}{2}} \sum_{i=1}^n \mathbf{g}_i(\theta_0) \right) \quad (2.55)$$

or in the notation of maximum likelihood,

$$\mathcal{J} \equiv \mathbb{E} \left[\frac{\partial l(\theta; Y_i)}{\partial \theta} \frac{\partial l(\theta; Y_i)}{\partial \theta'} \Big|_{\theta=\theta_0} \right]. \quad (2.56)$$

When the data are i.i.d., the scores or moment conditions should be i.i.d., and so the variance of the average is the average of the variance. The “plug-in” estimator for Σ uses the moment conditions evaluated at $\hat{\theta}$, and so the covariance estimator for method of moments applications with i.i.d. data is

$$\hat{\Sigma} = n^{-1} \sum_{i=1}^n \mathbf{g}_i(\hat{\theta}) \mathbf{g}_i(\hat{\theta})' \quad (2.57)$$

which is simply the average outer-product of the moment condition. The estimator of Σ in the maximum likelihood is identical replacing $\mathbf{g}_i(\hat{\theta})$ with $\partial l(\theta; y_i)/\partial \theta$ evaluated at $\hat{\theta}$,

$$\hat{\mathcal{J}} = n^{-1} \sum_{i=1}^n \frac{\partial l(\theta; y_i)}{\partial \theta} \frac{\partial l(\theta; y_i)}{\partial \theta'} \Big|_{\theta=\hat{\theta}}. \quad (2.58)$$

The “plug-in” estimator for the second derivative of the log-likelihood or the Jacobian of the moment conditions is similarly defined,

$$\hat{\mathbf{G}} = n^{-1} \sum_{i=1}^n \frac{\partial \mathbf{g}(\theta)}{\partial \theta'} \Big|_{\theta=\hat{\theta}} \quad (2.59)$$

or for maximum likelihood estimators

$$\hat{\mathcal{I}} = n^{-1} \sum_{i=1}^n -\frac{\partial^2 l(\theta; y_i)}{\partial \theta \partial \theta'} \Big|_{\theta=\hat{\theta}}. \quad (2.60)$$

2.4.6 Estimating Covariances with Dependent Data

The estimators in eq. (2.57) and eq. (2.58) are only appropriate when the moment conditions or scores are not correlated across i .²⁰ If the moment conditions or scores are correlated across observations the covariance estimator (but not the Jacobian estimator) must be changed to account for the dependence. Since Σ is defined as the variance of a sum it is necessary to account for both the sum of the variances *plus* all of the covariances.

$$\begin{aligned}\Sigma &\equiv \text{avar} \left(n^{-\frac{1}{2}} \sum_{i=1}^n \mathbf{g}_i(\boldsymbol{\theta}_0) \right) \\ &= \lim_{n \rightarrow \infty} n^{-1} \left(\sum_{i=1}^n \mathbb{E} [\mathbf{g}_i(\boldsymbol{\theta}_0) \mathbf{g}_i(\boldsymbol{\theta}_0)'] + \sum_{i=1}^{n-1} \sum_{j=i+1}^n \mathbb{E} [\mathbf{g}_j(\boldsymbol{\theta}_0) \mathbf{g}_{j-i}(\boldsymbol{\theta}_0)' + \mathbf{g}_{j-i}(\boldsymbol{\theta}_0) \mathbf{g}_j(\boldsymbol{\theta}_0)] \right)\end{aligned}\quad (2.61)$$

This expression depends on both the usual covariance of the moment conditions and on the covariance between the scores. When using i.i.d. data the second term vanishes since the moment conditions must be uncorrelated and so cross-products must have expectation 0.

If the moment conditions are correlated across i then covariance estimator must be adjusted to account for this. The obvious solution is to estimate the expectations of the cross terms in eq. (2.57) with their sample analogues, which would result in the covariance estimator

$$\hat{\Sigma}_{\text{DEP}} = n^{-1} \left[\sum_{i=1}^n \mathbf{g}_i(\hat{\boldsymbol{\theta}}) \mathbf{g}_i(\hat{\boldsymbol{\theta}})' + \sum_{i=1}^{n-1} \sum_{j=i+1}^n (\mathbf{g}_j(\hat{\boldsymbol{\theta}}) \mathbf{g}_{j-i}(\hat{\boldsymbol{\theta}})' + \mathbf{g}_{j-i}(\hat{\boldsymbol{\theta}}) \mathbf{g}_j(\hat{\boldsymbol{\theta}})') \right]. \quad (2.62)$$

This estimator is always zero since $\hat{\Sigma}_{\text{DEP}} = n^{-1} (\sum_{i=1}^n \mathbf{g}_i) (\sum_{i=1}^n \mathbf{g}_i)'$ and $\sum_{i=1}^n \mathbf{g}_i = \mathbf{0}$, and so $\hat{\Sigma}_{\text{DEP}}$ cannot be used in practice.²¹ One solution is to truncate the maximum lag to be something less than $n - 1$ (usually much less than $n - 1$), although the truncated estimator is not guaranteed to be positive definite. A better solution is to combine truncation with a weighting function (known as a *kernel*) to construct an estimator which will consistently estimate the covariance and is guaranteed to be positive definite. The most common covariance estimator of this type is the Newey and West (1987) covariance estimator. Covariance estimators for dependent data will be examined in more detail in the chapters on time-series data.

²⁰Since i.i.d. implies no correlation, the i.i.d. case is trivially covered.

²¹The scalar version of $\hat{\Sigma}_{\text{DEP}}$ may be easier to understand. If g_i is a scalar, then

$$\hat{\sigma}_{\text{DEP}}^2 = n^{-1} \left[\sum_{i=1}^n g_i^2(\hat{\boldsymbol{\theta}}) + 2 \sum_{i=1}^{n-1} \left(\sum_{j=i+1}^n g_j(\hat{\boldsymbol{\theta}}) g_{j-i}(\hat{\boldsymbol{\theta}}) \right) \right].$$

The first term is the usual variance estimator and the second term is the sum of the $(n - 1)$ covariance estimators. The more complicated expression in eq. (2.62) arises since order matters when multiplying vectors.

2.5 Hypothesis Testing

Econometrics models are estimated in order to test hypotheses, for example, whether a financial theory is supported by data or to determine if a model with estimated parameters can outperform a naïve forecast. Formal hypothesis testing begins by specifying the null hypothesis.

Definition 2.14 (Null Hypothesis). The null hypothesis, denoted H_0 , is a statement about the population values of some parameters to be tested. The null hypothesis is also known as the maintained hypothesis.

The null defines the condition on the population parameters that is to be tested. A null can be either simple, for example, $H_0 : \mu = 0$, or complex, which allows for testing of multiple hypotheses. For example, it is common to test whether data exhibit any predictability using a regression model

$$y_i = \theta_1 + \theta_2 x_{2,i} + \theta_3 x_{3,i} + \varepsilon_i, \quad (2.63)$$

and a composite null, $H_0 : \theta_2 = 0 \cap \theta_3 = 0$, often abbreviated $H_0 : \theta_2 = \theta_3 = 0$.²²

Null hypotheses cannot be accepted; the data can either lead to *rejection of the null* or a *failure to reject the null*. Neither option is “accepting the null”. The inability to accept the null arises since there are important cases where the data are not consistent with either the null or its testing complement, the alternative hypothesis.

Definition 2.15 (Alternative Hypothesis). The alternative hypothesis, denoted H_1 , is a complementary hypothesis to the null and determines the range of values of the population parameter that should lead to rejection of the null.

The alternative hypothesis specifies the population values of parameters for which the null should be rejected. In most situations, the alternative is the natural complement to the null in the sense that the null and alternative are exclusive of each other but inclusive of the range of the population parameter. For example, when testing whether a random variable has mean 0, the null is $H_0 : \mu = 0$ and the usual alternative is $H_1 : \mu \neq 0$.

In certain circumstances, usually motivated by theoretical considerations, one-sided alternatives are desirable. One-sided alternatives only reject for population parameter values on one side of zero and so test using one-sided alternatives may not reject even if both the null and alternative are false. Noting that a risk premium must be positive (if it exists), the null hypothesis of $H_0 : \mu = 0$ should be tested against the alternative $H_1 : \mu > 0$. This alternative indicates the null should only be rejected if there is compelling evidence that the mean is positive. These hypotheses further specify that data consistent with large negative values of μ should not lead to rejection. Focusing the alternative often leads to an increased probability to rejecting a false null. This occurs since the alternative is directed (positive values for μ), and less evidence is required to be convinced that the null is not valid.

Like null hypotheses, alternatives can be composite. The usual alternative to the null $H_0 : \theta_2 = 0 \cap \theta_3 = 0$ is $H_1 : \theta_2 \neq 0 \cup \theta_3 \neq 0$ and so the null should be rejected whenever any of the statements in the null are false – in other words if either or both $\theta_2 \neq 0$ or $\theta_3 \neq 0$. Alternatives can also be formulated as lists of exclusive outcomes.²³ When examining the relative precision of forecasting models, it is common to test the null that the forecast performance is equal against a composite alternative that the

²² \cap , the intersection operator, is used since the null requires both statements to be true.

²³ The \cup symbol indicates the union of the two alternatives.

forecasting performance is superior for model A or that the forecasting performance is superior for model B . If δ is defined as the average forecast performance difference, then the null is $H_0 : \delta = 0$ and the composite alternatives are $H_1^A : \delta > 0$ and $H_1^B : \delta < 0$, which indicate superior performance of models A and B , respectively.

Once the null and the alternative have been formulated, a hypothesis test is used to determine whether the data support the alternative.

Definition 2.16 (Hypothesis Test). A hypothesis test is a rule that specifies which values to reject H_0 in favor of H_1 .

Hypothesis testing requires a test statistic, for example, an appropriately standardized mean, and a critical value. The null is rejected when the test statistic is larger than the critical value.

Definition 2.17 (Critical Value). The critical value for a α -sized test, denoted C_α , is the value where a test statistic, T , indicates rejection of the null hypothesis when the null is true.

The region where the test statistic is outside of the critical value is known as the rejection region.

Definition 2.18 (Rejection Region). The rejection region is the region where $T > C_\alpha$.

An important event occurs when the null is correct but the hypothesis is rejected. This is known as a Type I error.

Definition 2.19 (Type I Error). A Type I error is the event that the null is rejected when the null is true.

A closely related concept is the size of the test. The size controls how often Type I errors should occur.

Definition 2.20 (Size). The size or level of a test, denoted α , is the probability of rejecting the null when the null is true. The size is also the probability of a Type I error.

Typical sizes include 1%, 5%, and 10%, although ideally, the selected size should reflect the decision makers preferences over incorrectly rejecting the null. When the opposite occurs, the null is not rejected when the alternative is true, a Type II error is made.

Definition 2.21 (Type II Error). A Type II error is the event that the null is not rejected when the alternative is true.

Type II errors are closely related to the power of a test.

Definition 2.22 (Power). The power of the test is the probability of rejecting the null when the alternative is true. The power is equivalently defined as 1 minus the probability of a Type II error.

The two error types, size and power are summarized in table 2.2.

A perfect test would have unit power against any alternative. In other words, whenever the alternative is true it would reject immediately. Practically the power of a test is a function of both the sample size and the distance between the population value of a parameter and its value under the null. A test is said to be consistent if the power of the test goes to 1 as $n \rightarrow \infty$ whenever the population value lies in the area defined by the alternative hypothesis. Consistency is an important characteristic of a

		Decision	
		Do not reject H_0	Reject H_0
Truth	H_0	Correct	Type I Error (Size)
	H_1	Type II Error	Correct (Power)

Table 2.2: Outcome matrix for a hypothesis test. The diagonal elements are both correct decisions. The off diagonal elements represent Type I error, when the null is rejected but is valid, and Type II error, when the null is not rejected and the alternative is true.

test, but it is usually considered more important to have correct size rather than to have high power. Because power can always be increased by distorting the size, and it is useful to consider a related measure known as the *size-adjusted power*. The size-adjusted power examines the power of a test in excess of size. Since a test should reject at size even when the null is true, it is useful to examine the percentage of times it *will* reject in excess of the percentage it *should* reject.

One useful tool for presenting results of test statistics is the p-value, or simply the p-val.

Definition 2.23 (P-value). The p-value is the probability of observing a value as large as the observed test statistic given the null is true. The p-value is also:

- The largest size (α) where the null hypothesis cannot be rejected.
- The smallest size where the null hypothesis can be rejected.

The primary advantage of a p-value is that it immediately demonstrates which test sizes would lead to rejection: anything above the p-value. It also improves on the common practice of reporting the test statistic alone since p-values can be interpreted without knowledge of the distribution of the test statistic. However, since it incorporates information about a specific test statistic and its associated distribution, the formula used to compute the p-value is problem specific.

A related representation is the confidence interval for a parameter.

Definition 2.24 (Confidence Interval). A confidence interval for a scalar parameter is the range of values, $\theta_0 \in (\underline{C}_\alpha, \bar{C}_\alpha)$ where the null $H_0 : \theta = \theta_0$ cannot be rejected for a size of α .

The formal definition of a confidence interval is not usually sufficient to uniquely identify the confidence interval. Suppose that a $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \sigma^2)$. The common 95% confidence interval is $(\hat{\theta} - 1.96\sigma, \hat{\theta} + 1.96\sigma)$. This set is known as the symmetric confidence interval and is formally defined as points $(\underline{C}_\alpha, \bar{C}_\alpha)$ where $\Pr(\theta_0) \in (\underline{C}_\alpha, \bar{C}_\alpha) = 1 - \alpha$ and $\underline{C}_\alpha - \theta = \theta - \bar{C}_\alpha$. An alternative, but still valid, confidence interval can be defined as $(-\infty, \hat{\theta} + 1.645\sigma^2)$. This would also contain the true value with probability 95%. In general, symmetric confidence intervals should be used, especially for asymptotically normal parameter estimates. In rare cases where symmetric confidence intervals are not appropriate, other options for defining a confidence interval include shortest interval, so that the confidence interval is defined as values $(\underline{C}_\alpha, \bar{C}_\alpha)$ where $\Pr(\theta_0) \in (\underline{C}_\alpha, \bar{C}_\alpha) = 1 - \alpha$ subject to

$\bar{C}_\alpha - \underline{C}_\alpha$ chosen to be as small as possible, or symmetric in probability, so that the confidence interval satisfies $\Pr(\theta_0) \in (\underline{C}_\alpha, \hat{\theta}) = \Pr(\theta_0) \in (\hat{\theta}, \bar{C}_\alpha) = 1/2 - \alpha/2$. When constructing confidence intervals for parameters that are asymptotically normal, these three definitions coincide.

2.5.0.1 Size and Power of a Test of the Mean with Normal Data

Suppose n i.i.d. normal random variables have unknown mean μ but known variance σ^2 and so the sample mean, $\bar{y} = n^{-1} \sum_{i=1}^n y_i$, is then distributed $N(\mu, \sigma^2/N)$. When testing a null that $H_0 : \mu = \mu_0$ against an alternative $H_1 : \mu \neq \mu_0$, the size of the test is the probability that the null is rejected when it is true. Since the distribution under the null is $N(\mu_0, \sigma^2/N)$ and the size can be set to α by selecting points where $\Pr(\hat{\mu} \in (\underline{C}_\alpha, \bar{C}_\alpha) | \mu = \mu_0) = 1 - \alpha$. Since the distribution is normal, one natural choice is to select the points symmetrically so that $\underline{C}_\alpha = \mu_0 + \frac{\sigma}{\sqrt{N}} \Phi^{-1}(\alpha/2)$ and $\bar{C}_\alpha = \mu_0 + \frac{\sigma}{\sqrt{N}} \Phi^{-1}(1 - \alpha/2)$ where $\Phi(\cdot)$ is the cdf of a standard normal.

The power of the test is defined as the probability the null is rejected when the alternative is true. This probability will depend on the population mean, μ_1 , the sample size, the test size and mean specified by the null hypothesis. When testing using an α -sized test, rejection will occur when $\hat{\mu} < \mu_0 + \frac{\sigma}{\sqrt{N}} \Phi^{-1}(\alpha/2)$ or $\hat{\mu} > \mu_0 + \frac{\sigma}{\sqrt{N}} \Phi^{-1}(1 - \alpha/2)$. Since under the alternative $\hat{\mu}$ is $N(\mu_1, \sigma^2)$, these probabilities will be

$$\Phi\left(\frac{\mu_0 + \frac{\sigma}{\sqrt{N}} \Phi^{-1}(\alpha/2) - \mu_1}{\frac{\sigma}{\sqrt{N}}}\right) = \Phi\left(\frac{\underline{C}_\alpha - \mu_1}{\frac{\sigma}{\sqrt{N}}}\right)$$

and

$$1 - \Phi\left(\frac{\mu_0 + \frac{\sigma}{\sqrt{N}} \Phi^{-1}(1 - \alpha/2) - \mu_1}{\frac{\sigma}{\sqrt{N}}}\right) = 1 - \Phi\left(\frac{\bar{C}_\alpha - \mu_1}{\frac{\sigma}{\sqrt{N}}}\right).$$

The total probability that the null is rejected is known as the power function,

$$\text{Power}(\mu_0, \mu_1, \sigma, \alpha, N) = \Phi\left(\frac{\underline{C}_\alpha - \mu_1}{\frac{\sigma}{\sqrt{N}}}\right) + 1 - \Phi\left(\frac{\bar{C}_\alpha - \mu_1}{\frac{\sigma}{\sqrt{N}}}\right).$$

A graphical illustration of the power is presented in figure 2.5. The null hypothesis is $H_0 : \mu = 0$ and the alternative distribution was drawn at $\mu_1 = .25$. The variance $\sigma^2 = 1$, $n = 5$, and the size was set to 5%. The highlighted regions indicate the power: the area under the alternative distribution, and hence the probability, which is outside of the critical values. The bottom panel illustrates the power curve for the same parameters allowing n to range from 5 to 1,000. When n is small, the power is low even for alternatives far from the null. As n grows the power increases and when $n = 1,000$, the power of the test is close to unity for alternatives greater than 0.1.

2.5.1 Statistical and Economic Significance

While testing can reject hypotheses and provide meaningful p-values, statistical significance is different from economic significance. Economic significance requires a more detailed look at the data than a simple hypothesis test. Establishing the statistical significance of a parameter is the first, and easy,

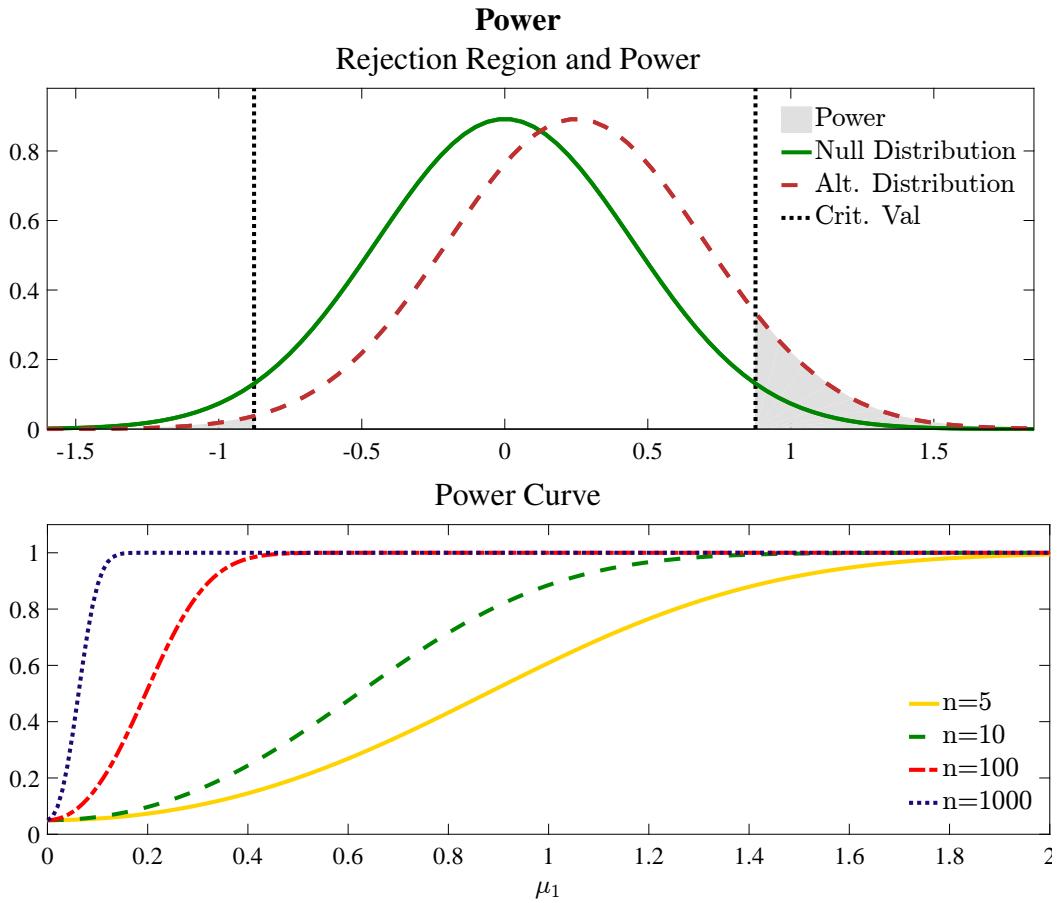


Figure 2.5: The top panel illustrates the power. The distribution of the mean under the null and alternative hypotheses were derived under that assumption that the data are i.i.d. normal with means $\mu_0 = 0$ and $\mu_1 = .25$, variance $\sigma^2 = 1$, $n = 5$ and $\alpha = .05$. The bottom panel illustrates the power function, in terms of the alternative mean, for the same parameters when $n = 5, 10, 100$ and $1,000$.

step. The more difficult step is to determine whether the effect is economically important. Consider a simple regression model

$$y_i = \theta_1 + \theta_2 x_{2,i} + \theta_3 x_{3,i} + \varepsilon_i \quad (2.64)$$

and suppose that the estimates of both θ_2 and θ_3 are statistically different from zero. This can happen for a variety of reasons, including having an economically small impact accompanied with a very large sample. To assess the relative contributions other statistics such as the percentage of the variation that can be explained by either variable alone and/or the range and variability of the xs.

Another important aspect of economic significance is that rejection of a hypothesis, while formally as a “yes” or “no” question, should be treated in a more continuous manner. The p-value of a test statistic is a useful tool in this regard that can provide a deeper insight into the strength of the rejection. A p-val of .00001 is not the same as a p-value of .09999 even though a 10% test would reject for either.

2.5.2 Specifying Hypotheses

Formalized in terms of θ , a null hypothesis is

$$H_0 : \mathbf{R}(\theta) = \mathbf{0} \quad (2.65)$$

where $\mathbf{R}(\cdot)$ is a function from \mathbb{R}^k to \mathbb{R}^m , $m \leq k$, where m represents the number of hypotheses in a composite null. While this specification of hypotheses is very flexible, testing non-linear hypotheses raises some subtle but important technicalities and further discussion will be reserved for later. Initially, a subset of all hypotheses, those in the linear equality restriction (LER) class, which can be specified as

$$H_0 : \mathbf{R}\theta - \mathbf{r} = \mathbf{0} \quad (2.66)$$

will be examined where \mathbf{R} is a m by k matrix and \mathbf{r} is a m by 1 vector. All hypotheses in the LER class can be written as weighted sums of model parameters,

$$\left[\begin{array}{l} R_{11}\theta_1 + R_{12}\theta_2 \dots + R_{1k}\theta_k = r_1 \\ R_{21}\theta_1 + R_{22}\theta_2 \dots + R_{2k}\theta_k = r_2 \\ \vdots \\ R_{m1}\theta_1 + R_{m2}\theta_2 \dots + R_{mk}\theta_k = r_i. \end{array} \right] \quad (2.67)$$

Each linear hypothesis is represented as a row in the above set of equations. Linear equality constraints can be used to test parameter restrictions on $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)'$ such as

$$\begin{aligned} \theta_1 &= 0 \\ 3\theta_2 + \theta_3 &= 1 \\ \sum_{j=1}^4 \theta_j &= 0 \\ \theta_1 &= \theta_2 = \theta_3 = 0. \end{aligned} \quad (2.68)$$

For example, the hypotheses in eq. (2.68) can be described in terms of \mathbf{R} and \mathbf{r} as

H_0	\mathbf{R}	\mathbf{r}
$\theta_1 = 0$	$[1 \ 0 \ 0 \ 0]$	0
$3\theta_2 + \theta_3 = 1$	$[0 \ 3 \ 1 \ 0]$	1
$\sum_{j=1}^4 \theta_j = 0$	$[1 \ 1 \ 1 \ 1]$	0
$\theta_1 = \theta_2 = \theta_3 = 0$	$[1 \ 0 \ 0 \ 0]$ $[0 \ 1 \ 0 \ 0]$ $[0 \ 0 \ 1 \ 0]$	$[0 \ 0 \ 0]'$

When using linear equality constraints, alternatives are generally formulated as $H_1 : \mathbf{R}\theta - \mathbf{r} \neq 0$. Once both the null and alternative hypotheses have been postulated, it is necessary to determine whether the data are consistent with the null hypothesis using one of the many tests.

2.5.3 The Classical Tests

Three classes of statistics will be described to test hypotheses: Wald, Lagrange Multiplier, and Likelihood Ratio. Wald tests are perhaps the most intuitive: they directly test whether $\mathbf{R}\hat{\theta} - \mathbf{r}$, the value under the null, is close to zero by exploiting the asymptotic normality of the estimated parameters. Lagrange Multiplier tests incorporate the constraint into the estimation problem using a Lagrangian. If the constraint has a small effect on the value of objective function, the Lagrange multipliers, often described as the shadow price of a constraint in an economic application, should be close to zero. The magnitude of the scores forms the basis of the LM test statistic. Finally, likelihood ratios test whether the data are less likely under the null than they are under the alternative. If these restrictions are not statistically meaningful, this ratio should be close to one since the difference in the log-likelihoods should be small.

2.5.4 Wald Tests

Wald test statistics are possibly the most natural method to test a hypothesis and are often the simplest to compute since only the unrestricted model must be estimated. Wald tests directly exploit the asymptotic normality of the estimated parameters to form test statistics with asymptotic χ_m^2 distributions. Recall that a χ_v^2 random variable is defined to be the sum of v independent standard normals squared, $\sum_{i=1}^v z_i^2$ where $z_i \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$. Recall that if \mathbf{z} is a m -dimension normal vector with mean μ and covariance Σ ,

$$\mathbf{z} \sim N(\mu, \Sigma) \quad (2.69)$$

then the standardized version of \mathbf{z} can be constructed as

$$\Sigma^{-\frac{1}{2}}(\mathbf{z} - \mu) \sim N(\mathbf{0}, \mathbf{I}). \quad (2.70)$$

Defining $\mathbf{w} = \Sigma^{-\frac{1}{2}}(\mathbf{z} - \mu) \sim N(\mathbf{0}, \mathbf{I})$, it is easy to see that $\mathbf{w}'\mathbf{w} = \sum_{m=1}^M w_m^2 \sim \chi_m^2$. In the usual case, the method of moments estimator, which nests ML and QML estimators as special cases, is asymptotically normal

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{G}^{-1}\Sigma(\mathbf{G}^{-1})'\right). \quad (2.71)$$

If null hypothesis, $H_0 : \mathbf{R}\theta = \mathbf{r}$ is true, it follows directly that

$$\sqrt{n}(\mathbf{R}\hat{\theta} - \mathbf{r}) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{R}\mathbf{G}^{-1}\Sigma(\mathbf{G}^{-1})'\mathbf{R}'\right). \quad (2.72)$$

This allows a test statistic to be formed

$$W = n(\mathbf{R}\hat{\theta} - \mathbf{r})' \left(\mathbf{R}\mathbf{G}^{-1}\Sigma(\mathbf{G}^{-1})'\mathbf{R}' \right)^{-1} (\mathbf{R}\hat{\theta} - \mathbf{r}) \quad (2.73)$$

which is the sum of the squares of m random variables, each asymptotically uncorrelated standard normal and so W is asymptotically χ_m^2 distributed. A hypothesis test with size α can be conducted by comparing W against $C_\alpha = F^{-1}(1 - \alpha)$ where $F(\cdot)$ is the cdf of a χ_m^2 . If $W \geq C_\alpha$ then the null is rejected.

There is one problem with the definition of W in eq. (2.73): it is infeasible since it depends on \mathbf{G} and Σ which are unknown. The usual practice is to replace the unknown elements of the covariance matrix with consistent estimates to compute a feasible Wald statistic,

$$W = n (\mathbf{R}\hat{\theta} - \mathbf{r})' (\mathbf{R}\hat{\mathbf{G}}^{-1}\hat{\Sigma}(\hat{\mathbf{G}}^{-1})' \mathbf{R}')^{-1} (\mathbf{R}\hat{\theta} - \mathbf{r}). \quad (2.74)$$

which has the same asymptotic distribution as the infeasible Wald test statistic.

2.5.4.1 t -tests

A t -test is a special case of a Wald and is applicable to tests involving a single hypothesis. Suppose the null is

$$H_0 : \mathbf{R}\theta = \mathbf{r}$$

where \mathbf{R} is 1 by k , and so

$$\sqrt{n} (\mathbf{R}\hat{\theta} - \mathbf{r}) \xrightarrow{d} N(\mathbf{0}, \mathbf{R}\mathbf{G}^{-1}\Sigma(\mathbf{G}^{-1})' \mathbf{R}').$$

The *studentized* version can be formed by subtracting the mean and dividing by the standard deviation,

$$t = \frac{\sqrt{n} (\mathbf{R}\hat{\theta} - \mathbf{r})}{\sqrt{\mathbf{R}\mathbf{G}^{-1}\Sigma(\mathbf{G}^{-1})' \mathbf{R}'}} \xrightarrow{d} N(0, 1). \quad (2.75)$$

and the test statistic can be compared to the critical values from a standard normal to conduct a hypothesis test. t -tests have an important advantage over the broader class of Wald tests – they can be used to test one-sided null hypotheses. A one-sided hypothesis takes the form $H_0 : \mathbf{R}\theta \geq \mathbf{r}$ or $H_0 : \mathbf{R}\theta \leq \mathbf{r}$ which are contrasted with one-sided alternatives of $H_1 : \mathbf{R}\theta < \mathbf{r}$ or $H_1 : \mathbf{R}\theta > \mathbf{r}$, respectively. When using a one-sided test, rejection occurs when $\mathbf{R}\theta - \mathbf{r}$ is statistically different from zero and when $\mathbf{R}\theta < \mathbf{r}$ or $\mathbf{R}\theta > \mathbf{r}$ as specified by the alternative.

t -tests are also used in commonly encountered test statistic, the t -stat, a test of the null that a parameter is 0 against an alternative that it is not. The t -stat is popular because most models are written in such a way that if a parameter $\theta = 0$ then it will have no impact.

Definition 2.25 (t -stat). The t -stat of a parameter θ_j is the t -test value of the null $H_0 : \theta_j = 0$ against a two-sided alternative $H_1 : \theta_j \neq 0$.

$$t\text{-stat} \equiv \frac{\hat{\theta}_j}{\sigma_{\hat{\theta}}} \quad (2.76)$$

where

$$\sigma_{\hat{\theta}} = \sqrt{\frac{\mathbf{e}_j \mathbf{G}^{-1} \Sigma (\mathbf{G}^{-1})' \mathbf{e}'_j}{n}} \quad (2.77)$$

and where \mathbf{e}_j is a vector of 0s with 1 in the j^{th} position.

Note that the t -stat is identical to the expression in eq. (2.75) when $\mathbf{R} = \mathbf{e}_j$ and $r = 0$. $\mathbf{R} = \mathbf{e}_j$ corresponds to a hypothesis test involving only element j of θ and $r = 0$ indicates that the null is $\theta_j = 0$.

A closely related measure is the standard error of a parameter. Standard errors are essentially standard deviations – square-roots of variance – except that the expression “standard error” is applied when describing the estimation error of a parameter while “standard deviation” is used when describing the variation in the data or population.

Definition 2.26 (Standard Error). The standard error of a parameter θ is the square root of the parameter’s variance,

$$\text{s.e.}(\hat{\theta}) = \sqrt{\sigma_{\hat{\theta}}^2} \quad (2.78)$$

where

$$\sigma_{\hat{\theta}}^2 = \frac{\mathbf{e}_j \mathbf{G}^{-1} \Sigma (\mathbf{G}^{-1})' \mathbf{e}'_j}{n} \quad (2.79)$$

and where \mathbf{e}_j is a vector of 0s with 1 in the j^{th} position.

2.5.5 Likelihood Ratio Tests

Likelihood ratio tests examine how “likely” the data are under the null and the alternative. If the hypothesis is valid then the data should be (approximately) equally likely under each. The LR test statistic is defined as

$$LR = -2(l(\tilde{\theta}; \mathbf{y}) - l(\hat{\theta}; \mathbf{y})) \quad (2.80)$$

where $\tilde{\theta}$ is defined

$$\tilde{\theta} = \arg \max_{\theta} l(\theta; \mathbf{y}) \quad (2.81)$$

subject to $\mathbf{R}\theta - \mathbf{r} = 0$

and $\hat{\theta}$ is the unconstrained estimator,

$$\hat{\theta} = \arg \max_{\theta} l(\theta; \mathbf{y}). \quad (2.82)$$

Under the null $H_0 : \mathbf{R}\theta - \mathbf{r} = 0$, the $LR \xrightarrow{d} \chi_m^2$. The intuition behind the asymptotic distribution of the LR can be seen in a second order Taylor expansion around parameters estimated under the null, $\tilde{\theta}$.

$$l(\mathbf{y}; \tilde{\theta}) = l(\mathbf{y}; \hat{\theta}) + (\tilde{\theta} - \hat{\theta})' \frac{\partial l(\mathbf{y}; \hat{\theta})}{\partial \theta} + \frac{1}{2} \sqrt{n} (\tilde{\theta} - \hat{\theta})' \frac{1}{n} \frac{\partial^2 l(\mathbf{y}; \hat{\theta})}{\partial \theta \partial \theta'} \sqrt{n} (\tilde{\theta} - \hat{\theta}) + R^3 \quad (2.83)$$

where R^3 is a remainder term that is vanishing as $n \rightarrow \infty$. Since $\hat{\theta}$ is an unconstrained estimator of θ_0 ,

$$\frac{\partial l(\mathbf{y}; \hat{\theta})}{\partial \theta} = \mathbf{0}$$

and

$$-2(l(\mathbf{y}; \tilde{\theta}) - l(\mathbf{y}; \hat{\theta})) \approx \sqrt{n}(\tilde{\theta} - \hat{\theta})' \left(-\frac{1}{n} \frac{\partial^2 l(\mathbf{y}; \hat{\theta})}{\partial \theta \partial \theta'} \right) \sqrt{n}(\tilde{\theta} - \hat{\theta}) \quad (2.84)$$

Under some mild regularity conditions, when the MLE is correctly specified

$$-\frac{1}{n} \frac{\partial^2 l(\mathbf{y}; \hat{\theta})}{\partial \theta \partial \theta'} \xrightarrow{P} -\mathbb{E}\left[\frac{\partial^2 l(\mathbf{y}; \theta_0)}{\partial \theta \partial \theta'}\right] = \mathcal{I},$$

and

$$\sqrt{n}(\tilde{\theta} - \hat{\theta}) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}^{-1}).$$

Thus,

$$\sqrt{n}(\tilde{\theta} - \hat{\theta})' \frac{1}{n} \frac{\partial^2 l(\mathbf{y}; \hat{\theta})}{\partial \theta \partial \theta'} \sqrt{n}(\tilde{\theta} - \hat{\theta}) \xrightarrow{d} \chi_m^2 \quad (2.85)$$

and so $2(l(\mathbf{y}; \hat{\theta}) - l(\mathbf{y}; \tilde{\theta})) \xrightarrow{d} \chi_m^2$. The only difficultly remaining is that the distribution of this quadratic form is a χ_m^2 an not a χ_k^2 since k is the dimension of the parameter vector. While formally establishing this is tedious, the intuition follows from the number of restrictions. If $\tilde{\theta}$ were unrestricted then it must be the case that $\tilde{\theta} = \hat{\theta}$ since $\hat{\theta}$ is defined as the unrestricted estimators. Applying a single restriction leave $k - 1$ free parameters in $\tilde{\theta}$ and thus it should be close to $\hat{\theta}$ except for this one restriction.

When models are correctly specified LR tests are very powerful against point alternatives (e.g. $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$). Another important advantage of the LR is that the covariance of the parameters does not need to be estimated. In many problems, accurate parameter covariances may be difficult to estimate, and imprecise covariance estimators produce adverse consequence for test statistics, such as size distortions where a 5% test will reject substantially more than 5% of the time when the null is true.

It is also important to note that the likelihood ratio *does not* have an asymptotic χ_m^2 when the assumed likelihood $f(\mathbf{y}; \theta)$ is misspecified. When this occurs the information matrix equality fails to hold and the asymptotic distribution of the LR is known as a *mixture of χ^2 distribution*. In practice, the assumed error distribution is often misspecified and so it is important that the distributional assumptions used to estimate θ are verified prior to using likelihood ratio tests.

Likelihood ratio tests are not available for method of moments estimators since no distribution function is assumed.²⁴

²⁴It is possible to construct a likelihood ratio-type statistic for method of moments estimators. Define

$$\mathbf{g}_n(\theta) = n^{-1} \sum_{i=1}^n \mathbf{g}_i(\theta)$$

2.5.6 Lagrange Multiplier, Score and Rao Tests

Lagrange Multiplier (LM), Score and Rao test are all the same statistic. While Lagrange Multiplier test may be the most appropriate description, describing the tests as score tests illustrates the simplicity of the test's construction. Score tests exploit the first order condition to test whether a null hypothesis is compatible with the data. Using the unconstrained estimator of θ , $\hat{\theta}$, the scores must be zero,

$$\frac{\partial l(\theta; \mathbf{y})}{\partial \theta} \Big|_{\theta=\hat{\theta}} = \mathbf{0}. \quad (2.86)$$

The score test examines whether the scores are “close” to zero – in a statistically meaningful way – when evaluated using the parameters estimated subject to the null restriction, $\tilde{\theta}$. Define

$$\mathbf{s}_i(\tilde{\theta}) = \frac{\partial l_i(\theta; y_i)}{\partial \theta} \Big|_{\theta=\tilde{\theta}} \quad (2.87)$$

as the i^{th} score, evaluated at the restricted estimator. If the null hypothesis is true, then

$$\sqrt{n} \left(n^{-1} \sum_{i=1}^n \mathbf{s}_i(\tilde{\theta}) \right) \xrightarrow{d} N(\mathbf{0}, \Sigma). \quad (2.88)$$

This forms the basis of the score test, which is computed as

$$LM = n \bar{\mathbf{s}}(\tilde{\theta})' \Sigma^{-1} \bar{\mathbf{s}}(\tilde{\theta}) \quad (2.89)$$

where $\bar{\mathbf{s}}(\tilde{\theta}) = n^{-1} \sum_{i=1}^n \mathbf{s}_i(\tilde{\theta})$. While this version is not feasible since it depends on Σ , the standard practice is to replace Σ with a consistent estimator and to compute the feasible score test,

$$LM = n \bar{\mathbf{s}}(\tilde{\theta})' \hat{\Sigma}^{-1} \bar{\mathbf{s}}(\tilde{\theta}) \quad (2.90)$$

where the estimator of Σ depends on the assumptions made about the scores. In the case where the scores are i.i.d. (usually because the data are i.i.d.),

$$\hat{\Sigma} = n^{-1} \sum_{i=1}^n \mathbf{s}_i(\tilde{\theta}) \mathbf{s}_i(\tilde{\theta})' \quad (2.91)$$

to be the average moment conditions evaluated at a parameter θ . The likelihood ratio-type statistic for method of moments estimators is defined as

$$\begin{aligned} LM &= n \mathbf{g}'_n(\tilde{\theta}) \hat{\Sigma}^{-1} \mathbf{g}_n(\tilde{\theta}) - n \mathbf{g}'_n(\hat{\theta}) \hat{\Sigma}^{-1} \mathbf{g}_n(\hat{\theta}) \\ &= n \mathbf{g}'_n(\tilde{\theta}) \hat{\Sigma}^{-1} \mathbf{g}_n(\tilde{\theta}) \end{aligned}$$

where the simplification is possible since $\mathbf{g}_n(\hat{\theta}) = 0$ and where

$$\hat{\Sigma} = n^{-1} \sum_{i=1}^n \mathbf{g}_i(\hat{\theta}) \mathbf{g}_i(\hat{\theta})'$$

is the sample covariance of the moment conditions evaluated at the *unrestricted* parameter estimates. This test statistic only differs from the LM test statistic in eq. (2.90) via the choice of the covariance estimator, and it should be similar in performance to the adjusted LM test statistic in eq. (2.92).

is a consistent estimator since $E[\mathbf{s}_i(\tilde{\theta})] = \mathbf{0}$ if the null is true. In practice a more powerful version of the LM test can be formed by subtracting the mean from the covariance estimator and using

$$\tilde{\Sigma} = n^{-1} \sum_{i=1}^n (\mathbf{s}_i(\tilde{\theta}) - \bar{\mathbf{s}}(\tilde{\theta})) (\mathbf{s}_i(\tilde{\theta}) - \bar{\mathbf{s}}(\tilde{\theta}))' \quad (2.92)$$

which must be smaller (in the matrix sense) than $\hat{\Sigma}$, although asymptotically, if the null is true, these two estimators will converge to the same limit. Like the Wald and the LR, the LM follows an asymptotic χ_m^2 distribution, and an LM test statistic will be rejected if $LM > C_\alpha$ where C_α is the $1 - \alpha$ quantile of a χ_m^2 distribution.

Scores test can be used with method of moments estimators by simply replacing the score of the likelihood with the moment conditions evaluated at the restricted parameter,

$$\mathbf{s}_i(\tilde{\theta}) = \mathbf{g}_i(\tilde{\theta}),$$

and then evaluating eq. (2.90) or (2.92).

2.5.7 Comparing and Choosing the Tests

All three of the classic tests, the Wald, likelihood ratio and Lagrange multiplier have the same limiting asymptotic distribution. In addition to all being asymptotically distributed as a χ_m^2 , they are all *asymptotically equivalent* in the sense they all have an identical asymptotic distribution and if one test rejects, the others will also reject. As a result, there is no asymptotic argument that one should be favored over the other.

The simplest justifications for choosing one over the others are practical considerations. Wald requires estimation *under the alternative* – the unrestricted model – and require an estimate of the asymptotic covariance of the parameters. LM tests require estimation *under the null* – the restricted model – and require an estimate of the asymptotic covariance of the scores evaluated at the restricted parameters. LR tests require both forms to be estimated but do not require any covariance estimates. On the other hand, Wald and LM tests can easily be made robust to many forms of misspecification by using the “sandwich” covariance estimator, $\mathbf{G}^{-1}\Sigma(\mathbf{G}^{-1})'$ for moment-based estimators or $\mathcal{I}^{-1}\mathcal{J}\mathcal{I}^{-1}$ for QML estimators. LR tests cannot be easily corrected and instead will have a non-standard distribution.

Models which are substantially easier to estimate under the null or alternative lead to a natural choice. If a model is easy to estimate in its restricted form, but not unrestricted LM tests are good choices. If estimation under the alternative is simpler then Wald tests are reasonable. If they are equally simple to estimate, and the distributional assumptions used in ML estimation are plausible, LR tests are likely the best choice. Empirically a relationship exists where $W \approx LR \geq LM$. LM is often smaller, and hence less likely to reject the null, since it estimates the covariance of the scores *under the null*. When the null may be restrictive, the scores will generally have higher variances when evaluated using the restricted parameters. The larger variances will lower the value of LM since the score covariance is inverted in the statistic. A simple method to correct this is to use the adjusted LM computed using the modified covariance estimator in eq. (2.92).

2.6 The Bootstrap and Monte Carlo

The bootstrap is an alternative technique for estimating parameter covariances and conducting inference. The name bootstrap is derived from the expression “to pick yourself up by your bootstraps” – a seemingly impossible task. The bootstrap, when initially proposed, was treated as an equally impossible feat, although it is now widely accepted as a valid, and in some cases, preferred method to plug-in type covariance estimation. The bootstrap is a simulation technique and is similar to Monte Carlo. However, unlike Monte Carlo, which requires a complete data-generating process, the bootstrap makes use of the observed data to simulate the data – hence the similarity to the original turn-of-phrase.

Monte Carlo is an integration technique that uses simulation to approximate the underlying distribution of the data. Suppose $Y_i \stackrel{\text{i.i.d.}}{\sim} F(\theta)$ where F is some distribution, and that interest is in the $E[g(Y)]$. Further suppose it is possible to simulate from $F(\theta)$ so that a sample $\{y_i\}$ can be constructed. Then

$$n^{-1} \sum_{i=1}^n g(Y_i) \xrightarrow{P} E[g(Y)]$$

as long as this expectation exists since the simulated data are i.i.d. by construction.

The observed data can be used to compute the empirical cdf.

Definition 2.27 (Empirical cdf). The empirical cdf is defined

$$\hat{F}(c) = n^{-1} \sum_{i=1}^n I_{[y_i < c]}.$$

As long as \hat{F} is close to F , then the empirical cdf can be used to simulate random variables which should be approximately distributed F , and simulated data from the empirical cdf should have similar statistical properties (mean, variance, etc.) as data simulated from the true population cdf. The empirical cdf is a coarse step function and so *only* values which have been observed can be simulated, and so simulating from the empirical cdf of the data is identical to re-sampling the original data. In other words, the observed data can be directly used to simulate the from the underlying (unknown) cdf.

Figure 2.6 shows the population cdf for a standard normal and two empirical cdfs, one estimated using $n = 20$ observations and the other using $n = 1,000$. The coarse empirical cdf highlights the stair-like features of the empirical cdf estimate which restrict random numbers generated using the empirical cdf to coincide with the data used to compute the empirical cdf.

The bootstrap can be used for a variety of purposes. The most application of a bootstrap is to estimate the covariance matrix of some estimated parameters. This is an alternative to the usual plug-in type estimator and is simple to implement when the estimator is available in closed form.

Algorithm 2.1 (i.i.d. Nonparametric Bootstrap Covariance).

1. Generate a set of n uniform integers $\{j_i\}_{i=1}^n$ on $[1, 2, \dots, n]$.
2. Construct a simulated sample $\{y_{j_i}\}$.
3. Estimate the parameters of interest using $\{y_{j_i}\}$, and denote the estimate $\tilde{\theta}_b$.

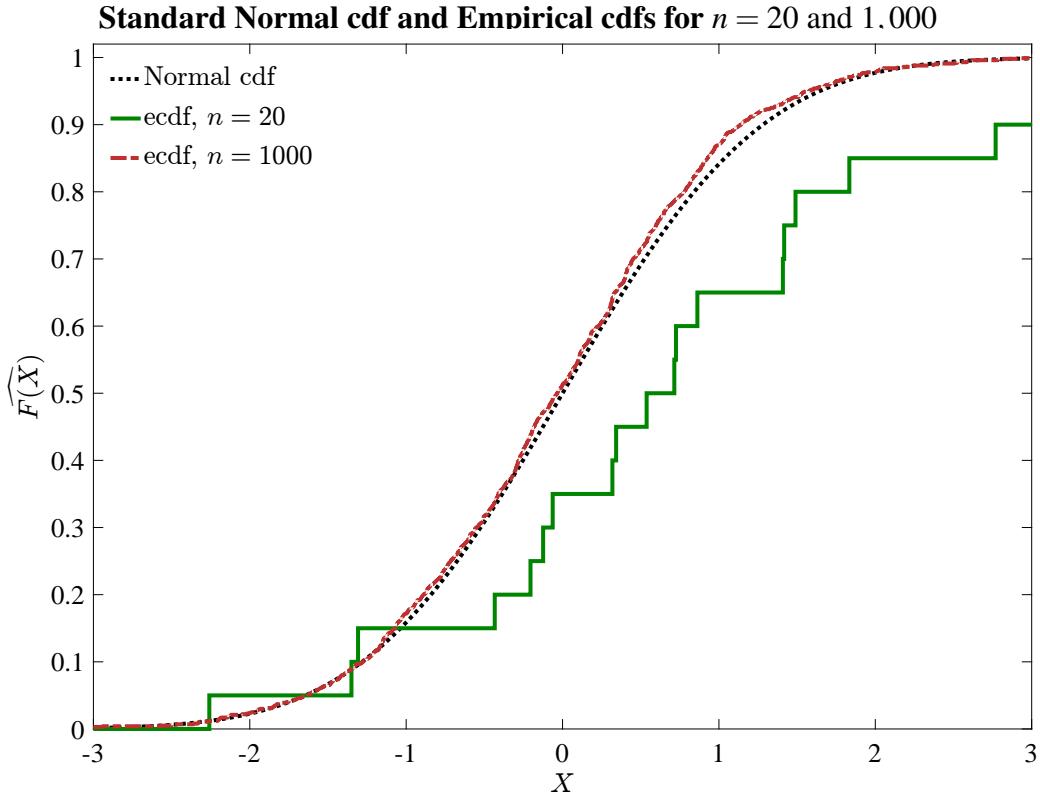


Figure 2.6: These three lines represent the population cdf of a standard normal, and two empirical cdfs constructed from simulated data. The very coarse empirical cdf is based on 20 observations and clearly highlights the step-nature of empirical cdfs. The other empirical cdf, which is based on 1,000 observations, appears smoother but is still a step function.

4. Repeat steps 1 through 3 a total of B times.

5. Estimate the variance of $\hat{\theta}$ using

$$\widehat{\mathbf{V}}[\hat{\theta}] = B^{-1} \sum_{b=1}^B (\tilde{\theta}_j - \hat{\theta}) (\tilde{\theta}_j - \hat{\theta})'$$

or alternatively

$$\widehat{\mathbf{V}}[\hat{\theta}] = B^{-1} \sum_{b=1}^B (\tilde{\theta}_j - \bar{\theta}) (\tilde{\theta}_j - \bar{\theta})'$$

The variance estimator that comes from this algorithm cannot be directly compared to the asymptotic covariance estimator since the bootstrap covariance is converging to 0. Normalizing the bootstrap covariance estimate by \sqrt{n} will allow comparisons and direct application of the test statistics based on the asymptotic covariance. Note that when using a conditional model, the vector $[y_i \mathbf{x}_i']'$ should be jointly bootstrapped. Aside from this small modification to step 2, the remainder of the procedure remains valid.

The nonparametric bootstrap is closely related to the residual bootstrap, at least when it is possible to appropriately define a residual. For example, when $Y_i|\mathbf{X}_i \sim N(\beta' \mathbf{x}_i, \sigma^2)$, the residual can be defined $\hat{\epsilon}_i = y_i - \hat{\beta}' \mathbf{x}_i$. Alternatively if $Y_i|\mathbf{X}_i \sim \text{Scaled-}\chi^2_v(\exp(\beta' \mathbf{x}_i))$, then $\hat{\epsilon}_i = y_i / \sqrt{\hat{\beta}' \mathbf{x}}$. The residual bootstrap can be used whenever it is possible to express $y_i = g(\theta, \epsilon_i, \mathbf{x}_i)$ for some known function g .

Algorithm 2.2 (i.i.d. Residual Bootstrap Covariance).

1. Generate a set of n uniform integers $\{j_i\}_{i=1}^n$ on $[1, 2, \dots, n]$.
2. Construct a simulated sample $\{\hat{\epsilon}_{j_i}, \mathbf{x}_{j_i}\}$ and define $\tilde{y}_i = g(\hat{\theta}, \tilde{\epsilon}_i, \tilde{\mathbf{x}}_i)$ where $\tilde{\epsilon}_i = \hat{\epsilon}_{j_i}$ and $\tilde{\mathbf{x}}_i = \mathbf{x}_{j_i}$.²⁵
3. Estimate the parameters of interest using $\{\tilde{y}_i, \tilde{\mathbf{x}}_i\}$, and denote the estimate $\tilde{\theta}_b$.
4. Repeat steps 1 through 3 a total of B times.
5. Estimate the variance of $\hat{\theta}$ using

$$\widehat{\mathbb{V}}[\hat{\theta}] = B^{-1} \sum_{b=1}^B (\tilde{\theta}_b - \hat{\theta}) (\tilde{\theta}_b - \hat{\theta})'$$

or alternatively

$$\widehat{\mathbb{V}}[\hat{\theta}] = B^{-1} \sum_{b=1}^B (\tilde{\theta}_b - \bar{\theta}) (\tilde{\theta}_b - \bar{\theta})'$$

It is important to emphasize that the bootstrap is not, generally, a better estimator of parameter covariance than standard plug-in estimators.²⁶ Asymptotically both are consistent and can be used equivalently. Additionally, i.i.d. bootstraps can only be applied to (conditionally) i.i.d. data and using an inappropriate bootstrap will produce an inconsistent estimator. When data have dependence it is necessary to use an alternative bootstrap scheme.

When the interest lies in confidence intervals, an alternative procedure that directly uses the empirical quantiles of the bootstrap parameter estimates can be constructed (known as the percentile method).

Algorithm 2.3 (i.i.d. Nonparametric Bootstrap Confidence Interval).

1. Generate a set of n uniform integers $\{j_i\}_{i=1}^n$ on $[1, 2, \dots, n]$.
2. Construct a simulated sample $\{y_{j_i}\}$.
3. Estimate the parameters of interest using $\{y_{j_i}\}$, and denote the estimate $\tilde{\theta}_b$.
4. Repeat steps 1 through 3 a total of B times.

²⁵In some models, it is possible to use independent indices on $\hat{\epsilon}$ and \mathbf{x} , such as in a linear regression when the data are conditionally homoskedastic (See chapter 3). In general it is not possible to explicitly break the link between ϵ_i and \mathbf{x}_i , and so these should usually be resampled using the same indices.

²⁶There are some problem-dependent bootstraps that are more accurate than plug-in estimators in an asymptotic sense. These are rarely encountered in financial economic applications.

5. Estimate the $1 - \alpha$ confidence interval of $\hat{\theta}_k$ using

$$[q_{\alpha/2}(\{\tilde{\theta}_k\}), q_{1-\alpha/2}(\{\tilde{\theta}_k\})]$$

where $q_\alpha(\{\tilde{\theta}_k\})$ is the empirical α quantile of the bootstrap estimates. 1-sided lower confidence intervals can be constructed as

$$\left[\underline{R}(\theta_k), q_{1-\alpha}(\{\tilde{\theta}_k\})\right]$$

and 1-sided upper confidence intervals can be constructed as

$$\left[q_\alpha(\{\tilde{\theta}_k\}), \overline{R}(\theta_k)\right]$$

where $\underline{R}(\theta_k)$ and $\overline{R}(\theta_k)$ are the lower and upper extremes of the range of θ_k (possibly $\pm\infty$).

The percentile method can also be used directly to compute P-values of test statistics. This requires enforcing the null hypothesis on the data and so is somewhat more involved. For example, suppose the null hypothesis is $E[y_i] = 0$. This can be enforced by replacing the original data with $\tilde{y}_i = y_i - \bar{y}$ in step 2 of the algorithm.

Algorithm 2.4 (i.i.d. Nonparametric Bootstrap P-value).

1. Generate a set of n uniform integers $\{j_i\}_{i=1}^n$ on $[1, 2, \dots, n]$.
2. Construct a simulated sample using data where the null hypothesis is true, $\{\tilde{y}_{j_i}\}$.
3. Compute the test statistic of interest using $\{\tilde{y}_{j_i}\}$, and denote the statistic $T(\tilde{\theta}_b)$.
4. Repeat steps 1 through 3 a total of B times.
5. Compute the bootstrap P-value using

$$\widehat{P-val} = B^{-1} \sum_{b=1}^B I_{[T(\hat{\theta}) \leq T(\tilde{\theta}_b)]}$$

for 1-sided tests where the rejection region is for large values (e.g. a Wald test). When using 2-sided tests, compute the bootstrap P-value using

$$\widehat{P-val} = B^{-1} \sum_{b=1}^B I_{[|T(\hat{\theta})| \leq |T(\tilde{\theta}_b)|]}$$

The test statistic may depend on a covariance matrix. When this is the case, the covariance matrix is usually estimated from the bootstrapped data using a plug-in method. Alternatively, it is possible to use any other consistent estimator (when the null is true) of the asymptotic covariance, such as one based on an initial (separate) bootstrap.

When models are maximum likelihood based, so that a complete model for the data is specified, it is possible to use a parametric form of the bootstrap to estimate covariance matrices. This procedure is virtually identical to standard Monte Carlo except that the initial estimate $\hat{\theta}$ is used in the simulation.

Algorithm 2.5 (i.i.d. Parametric Bootstrap Covariance (Monte Carlo)).

1. Simulate a set of n i.i.d. draws $\{\tilde{y}_i\}$ from $F(\hat{\theta})$.
2. Estimate the parameters of interest using $\{\tilde{y}_i\}$, and denote the estimates $\tilde{\theta}_b$.
3. Repeat steps 1 through 4 a total of B times.
4. Estimate the variance of $\hat{\theta}$ using

$$\text{V}[\hat{\theta}] = B^{-1} \sum_{b=1}^B (\tilde{\theta}_b - \hat{\theta})(\tilde{\theta}_b - \hat{\theta})'$$

or alternatively

$$\text{V}[\hat{\theta}] = B^{-1} \sum_{b=1}^B (\tilde{\theta}_b - \bar{\theta})(\tilde{\theta}_b - \bar{\theta})'$$

When models use conditional maximum likelihood, it is possible to use parametric bootstrap as part of a two-step procedure. First, apply a nonparametric bootstrap to the conditioning data $\{\mathbf{x}_i\}$, and then, using the bootstrapped conditioning data, simulate $Y_i \sim F(\hat{\theta}|\tilde{\mathbf{X}}_i)$. This is closely related to the residual bootstrap, only the assumed parametric distribution F is used in place of the data-derived residuals.

2.7 Inference on Financial Data

Inference will be covered in greater detail in conjunction with specific estimators and models, such as linear regression or ARCH models. These examples examine relatively simple hypotheses to illustrate the steps required in testing hypotheses.

2.7.1 Testing the Market Premium

Testing the market premium is a cottage industry. While current research is more interested in predicting the market premium, testing whether the market premium is significantly different from zero is a natural application of the tools introduced in this chapter. Let λ denote the market premium and let σ^2 be the variance of the return. Since the market is a traded asset it must be the case that the premium for holding market risk is the same as the mean of the market return. Monthly data for the Value Weighted Market (VWM) and the risk-free rate (R_f) was available between January 1927 and June 2008. Data for the VWM was drawn from CRSP and data for the risk-free rate was available from Ken French's data library. Excess returns on the market are defined as the return to holding the market minus the risk-free rate, $VWM_i^e = VWM_i - R_f$. The excess returns along with a kernel density plot are presented in figure 2.7. Excess returns are both negatively skewed and heavy-tailed – October 1987 is 5 standard deviations from the mean.

The mean and variance can be computed using the method of moments as detailed in section 2.1.4, and the covariance of the mean and the variance can be computed using the estimators described in

section 2.4.1. The estimates were calculated according to

$$\begin{bmatrix} \hat{\lambda} \\ \hat{\sigma}^2 \end{bmatrix} = \begin{bmatrix} n^{-1} \sum_{i=1}^n VWM_i^e \\ n^{-1} \sum_{i=1}^n (VWM_i^e - \hat{\lambda})^2 \end{bmatrix}$$

and, defining $\hat{\varepsilon}_i = VWM_i^e - \hat{\lambda}$, the covariance of the moment conditions was estimated by

$$\hat{\Sigma} = n^{-1} \begin{bmatrix} \sum_{i=1}^n \hat{\varepsilon}_i^2 & \sum_{i=1}^n \hat{\varepsilon}_i (\hat{\varepsilon}_i^2 - \hat{\sigma}^2) \\ \sum_{i=1}^n \hat{\varepsilon}_i (\hat{\varepsilon}_i^2 - \hat{\sigma}^2) & \sum_{i=1}^n (\hat{\varepsilon}_i^2 - \hat{\sigma}^2)^2 \end{bmatrix}.$$

Since the plim of the Jacobian is $-\mathbf{I}_2$, the parameter covariance is also $\hat{\Sigma}$. Combining these two results with a Central Limit Theorem (assumed to hold), the asymptotic distribution is

$$\sqrt{n} [\theta - \hat{\theta}] \xrightarrow{d} N(\mathbf{0}, \Sigma)$$

where $\theta = (\lambda, \sigma^2)'$. These produce the results in the first two rows of table 2.3.

These estimates can also be used to make inference on the standard deviation, $\sigma = \sqrt{\sigma^2}$ and the Sharpe ratio, $S = \lambda/\sigma$. The derivation of the asymptotic distribution of the Sharpe ratio was presented in 2.4.4.1 and the asymptotic distribution of the standard deviation can be determined in a similar manner where $\mathbf{d}(\theta) = \sqrt{\sigma^2}$ and so

$$\mathbf{D}(\theta) = \frac{\partial \mathbf{d}(\theta)}{\partial \theta'} = \begin{bmatrix} 0 & \frac{1}{2\sqrt{\sigma^2}} \end{bmatrix}.$$

Combining this expression with the asymptotic distribution for the estimated mean and variance, the asymptotic distribution of the standard deviation estimate is

$$\sqrt{n} (\hat{\sigma} - \sigma) \xrightarrow{d} N\left(0, \frac{\mu_4 - \sigma^4}{4\sigma^2}\right).$$

which was computed by dividing the [2,2] element of the parameter covariance by $4\hat{\sigma}^2$.

2.7.1.1 Bootstrap Implementation

The bootstrap can be used to estimate parameter covariance, construct confidence intervals – either used the estimated covariance or the percentile method, and to tabulate the P-value of a test statistic. Estimating the parameter covariance is simple – the data is resampled to create a simulated sample with n observations and the mean and variance are estimated. This is repeated 10,000 times and the parameter covariance is estimated using

$$\begin{aligned} \hat{\Sigma} &= B^{-1} \sum_{b=1}^B \left(\begin{bmatrix} \tilde{\mu}_b \\ \tilde{\sigma}_b^2 \end{bmatrix} - \begin{bmatrix} \hat{\mu}_b \\ \hat{\sigma}_b^2 \end{bmatrix} \right) \left(\begin{bmatrix} \tilde{\mu}_b \\ \tilde{\sigma}_b^2 \end{bmatrix} - \begin{bmatrix} \hat{\mu}_b \\ \hat{\sigma}_b^2 \end{bmatrix} \right)' \\ &= B^{-1} \sum_{b=1}^B (\tilde{\theta}_b - \hat{\theta}) (\tilde{\theta}_b - \hat{\theta})'. \end{aligned}$$

The percentile method can be used to construct confidence intervals for the parameters as estimated and for functions of parameters such as the Sharpe ratio. Constructing the confidence intervals

Parameter	Estimate	Standard Error	t-stat
λ	0.627	0.173	3.613
σ^2	29.41	2.957	9.946
σ	5.423	0.545	9.946
$\frac{\lambda}{\sigma}$	0.116	0.032	3.600

Table 2.3: Parameter estimates and standard errors for the market premium (λ), the variance of the excess return (σ^2), the standard deviation of the excess return (σ) and the Sharpe ratio ($\frac{\lambda}{\sigma}$). Estimates and variances were computed using the method of moments. The standard errors for σ and $\frac{\lambda}{\sigma}$ were computed using the delta method.

Parameter	Estimate	Standard Error	Bootstrap	Confidence Interval
			Lower	Upper
λ	0.627	0.174	0.284	0.961
σ^2	29.41	2.964	24.04	35.70
σ	5.423	0.547	4.903	5.975
$\frac{\lambda}{\sigma}$	0.116	0.032	0.052	0.179

$H_0 : \lambda = 0$	
P-value	3.00×10^{-4}

Table 2.4: Parameter estimates, bootstrap standard errors and confidence intervals (based on the percentile method) for the market premium (λ), the variance of the excess return (σ^2), the standard deviation of the excess return (σ) and the Sharpe ratio ($\frac{\lambda}{\sigma}$). Estimates were computed using the method of moments. The standard errors for σ and $\frac{\lambda}{\sigma}$ were computed using the delta method using the bootstrap covariance estimator.

for a function of the parameters requires constructing the function of the estimated parameters using each simulated sample and then computing the confidence interval using the empirical quantile of these estimates. Finally, the test P-value for the statistic for the null $H_0 : \lambda = 0$ can be computed directly by transforming the returns so that they have mean 0 using $\tilde{r}_i = r_i - \bar{r}_i$. The P-value can be tabulated using

$$\widehat{P\text{-val}} = B^{-1} \sum_{b=1}^B I_{[\bar{r} \leq \bar{r}_b]}$$

where \bar{r}_b is the average from bootstrap replication b . Table 2.4 contains the bootstrap standard errors, confidence intervals based on the percentile method and the bootstrap P-value for testing whether the mean return is 0. The standard errors are virtually identical to those estimated using the plug-in method, and the confidence intervals are similar to $\hat{\theta}_k \pm 1.96s.e.(\theta_k)$. The null that the average return is 0 is also strongly rejected.

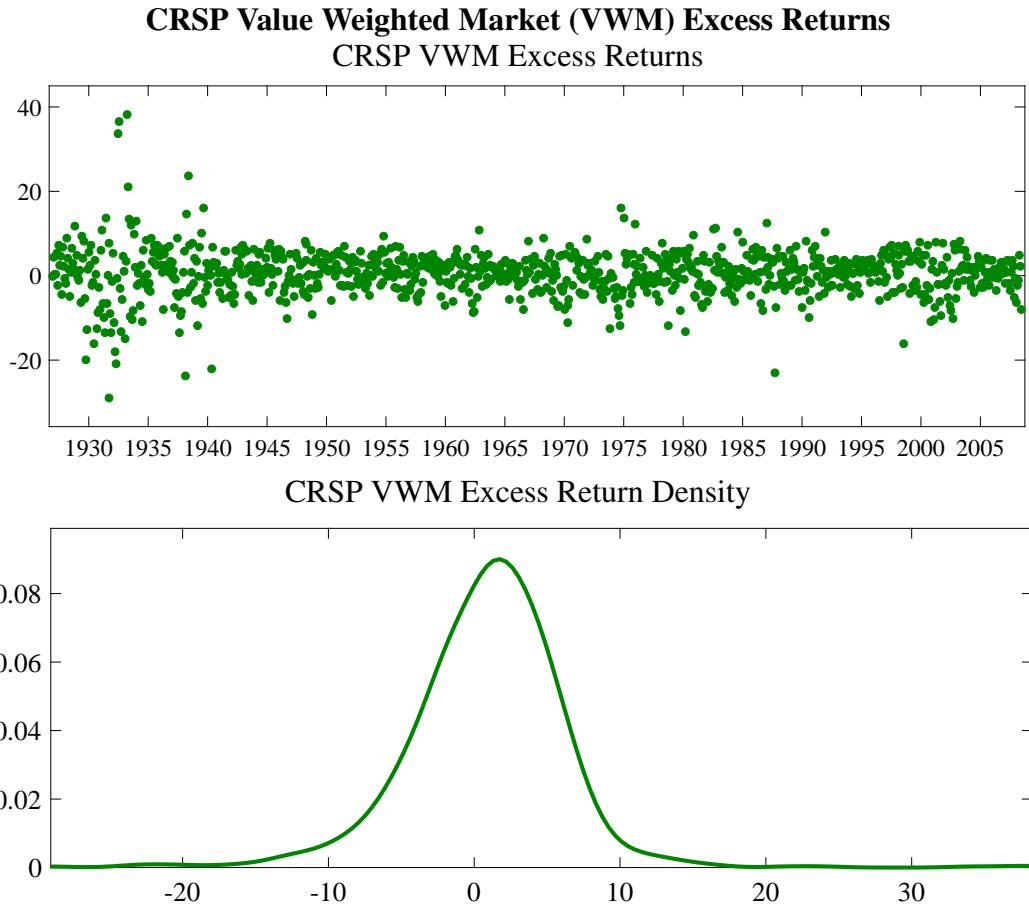


Figure 2.7: These two plots contain the returns on the VWM (top panel) in excess of the risk free rate and a kernel estimate of the density (bottom panel). While the mode of the density (highest peak) appears to be clearly positive, excess returns exhibit strong negative skew and are heavy tailed.

2.7.2 Is the NASDAQ Riskier than the S&P 100?

A second application examines the riskiness of the NASDAQ and the S&P 100. Both of these indices are value-weighted and contain 100 companies. The NASDAQ 100 contains only companies that trade on the NASDAQ while the S&P 100 contains large companies that trade on either the NYSE or the NASDAQ.

The null hypothesis is that the variances are the same, $H_0 : \sigma_{SP}^2 = \sigma_{ND}^2$, and the alternative is that the variance of the NASDAQ is larger, $H_1 : \sigma_{ND}^2 > \sigma_{SP}^2$.²⁷ The null and alternative can be reformulated as a test that $\delta = \sigma_{ND}^2 - \sigma_{SP}^2$ is equal to zero against an alternative that it is greater than zero. The estimation of the parameters can be formulated as a method of moments problem,

²⁷It may also be interesting to test against a two-sided alternative that the variances are unequal, $H_1 : \sigma_{ND}^2 \neq \sigma_{SP}^2$.

Parameter	Estimate	Daily Data				
		Std. Error/Correlation				
μ_{SP}	9.06	3.462	-0.274	0.767	-0.093	
σ_{SP}	17.32	-0.274	0.709	-0.135	0.528	
μ_{ND}	9.73	0.767	-0.135	4.246	-0.074	
σ_{NS}	21.24	-0.093	0.528	-0.074	0.443	
Test Statistics						
δ	0.60	$\hat{\sigma}_\delta$	0.09	<i>t</i> -stat	6.98	
Monthly Data						
Parameter	Estimate	Std. Error/Correlation				
		3.022	-0.387	0.825	-0.410	
μ_{SP}	8.61	-0.387	1.029	-0.387	0.773	
σ_{SP}	15.11	0.825	-0.387	4.608	-0.418	
μ_{ND}	9.06	-0.410	0.773	-0.418	1.527	
σ_{NS}	23.04					
Test Statistics						
δ	25.22	$\hat{\sigma}_\delta$	4.20	<i>t</i> -stat	6.01	

Table 2.5: Estimates, standard errors and correlation matrices for the S&P 100 and NASDAQ 100. The top panel uses daily return data between January 3, 1983, and December 31, 2007 (6,307 days) to estimate the parameter values in the left-most column. The rightmost 4 columns contain the parameter standard errors (diagonal elements) and the parameter correlations (off-diagonal elements). The bottom panel contains estimates, standard errors, and correlations from monthly data between January 1983 and December 2007 (300 months). Parameter and covariance estimates have been annualized. The test statistics (and related quantities) were performed and reported on the original (non-annualized) values.

$$\begin{bmatrix} \hat{\mu}_{SP} \\ \hat{\sigma}_{SP}^2 \\ \hat{\mu}_{ND} \\ \hat{\sigma}_{ND}^2 \end{bmatrix} = n^{-1} \sum_{i=1}^n \begin{bmatrix} r_{SP,i} \\ (r_{SP,i} - \hat{\mu}_{SP})^2 \\ r_{ND,i} \\ (r_{ND,i} - \hat{\mu}_{ND})^2 \end{bmatrix}$$

Inference can be performed by forming the moment vector using the estimated parameters, \mathbf{g}_i ,

$$\mathbf{g}_i = \begin{bmatrix} r_{SP,i} - \mu_{SP} \\ (r_{SP,i} - \mu_{SP})^2 - \sigma_{SP}^2 \\ r_{ND,i} - \mu_{ND} \\ (r_{ND,i} - \mu_{ND})^2 - \sigma_{ND}^2 \end{bmatrix}$$

and recalling that the asymptotic distribution is given by

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\left(0, \mathbf{G}^{-1}\Sigma(\mathbf{G}')^{-1}\right).$$

Using the set of moment conditions,

$$\begin{aligned} \mathbf{G} &= \text{plim}_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \begin{bmatrix} -1 & 0 & 0 & 0 \\ -2(r_{SP,i} - \mu_{SP}) & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & -2(r_{ND,i} - \mu_{ND}) & -1 \end{bmatrix} \\ &= -\mathbf{I}_4. \end{aligned}$$

Σ can be estimated using the moment conditions evaluated at the estimated parameters, $\mathbf{g}_i(\hat{\theta})$,

$$\hat{\Sigma} = n^{-1} \sum_{i=1}^n \mathbf{g}_i(\hat{\theta}) \mathbf{g}_i'(\hat{\theta}).$$

Noting that the (2,2) element of Σ is the variance of $\hat{\sigma}_{SP}^2$, the (4,4) element of Σ is the variance of $\hat{\sigma}_{ND}^2$ and the (2,4) element is the covariance of the two, the variance of $\hat{\delta} = \hat{\sigma}_{ND}^2 - \hat{\sigma}_{SP}^2$ can be computed as the sum of the variances minus two times the covariance, $\Sigma_{[2,2]} + \Sigma_{[4,4]} - 2\Sigma_{[2,4]}$. Finally a *one-sided* *t*-test can be performed to test the null.

Data was taken from Yahoo! finance between January 1983 and December 2008 at both the daily and monthly frequencies. Parameter estimates are presented in table 2.5. The table also contains the parameter standard errors – the square-root of the asymptotic covariance divided by the number of observations ($\sqrt{\Sigma_{[i,i]}/n}$) – along the diagonal and the parameter correlations – $\Sigma_{[i,j]}/\sqrt{\Sigma_{[i,i]}\Sigma_{[j,j]}}$ – in the off-diagonal positions. The top panel contains results for daily data while the bottom contains results for monthly data. Returns scaled by 100 were used in both panels .

All parameter estimates are reported in annualized form, which requires multiplying daily (monthly) mean estimates by 252 (12), and daily (monthly) volatility estimated by $\sqrt{252}$ ($\sqrt{12}$). Additionally, the delta method was used to adjust the standard errors on the volatility estimates since the actual parameter estimates were the means and variances. Thus, the reported parameter variance covariance matrix has the form

$$\mathbf{D}(\hat{\theta}) \hat{\Sigma} \mathbf{D}(\hat{\theta}) = \begin{bmatrix} 252 & 0 & 0 & 0 \\ 0 & \frac{\sqrt{252}}{2\sigma_{SP}} & 0 & 0 \\ 0 & 0 & 252 & 0 \\ 0 & 0 & 0 & \frac{\sqrt{252}}{2\sigma_{ND}} \end{bmatrix} \hat{\Sigma} \begin{bmatrix} 252 & 0 & 0 & 0 \\ 0 & \frac{\sqrt{252}}{2\sigma_{SP}} & 0 & 0 \\ 0 & 0 & 252 & 0 \\ 0 & 0 & 0 & \frac{\sqrt{252}}{2\sigma_{ND}} \end{bmatrix}.$$

In both cases δ is positive with a *t*-stat greater than 6, indicating a strong rejection of the null in favor of the alternative. Since this was a one-sided test, the 95% critical value would be 1.645 ($\Phi(.95)$).

This test could also have been implemented using an LM test, which requires estimating the two mean parameters but restricting the variances to be equal. One $\tilde{\theta}$ is estimated, the LM test statistic is computed as

$$LM = n\mathbf{g}_n(\tilde{\theta}) \hat{\Sigma}^{-1} \mathbf{g}_n'(\tilde{\theta})$$

Parameter	Estimate	Daily Data			
		Boot	Std.	Error	Correlation
μ_{SP}	9.06	3.471	-0.276	0.767	-0.097
σ_{SP}	17.32	-0.276	0.705	-0.139	0.528
μ_{ND}	9.73	0.767	-0.139	4.244	-0.079
σ_{NS}	21.24	-0.097	0.528	-0.079	0.441

Parameter	Estimate	Monthly Data			
		Bootstrap	Std.	Error	Correlation
μ_{SP}	8.61	3.040	-0.386	0.833	-0.417
σ_{SP}	15.11	-0.386	1.024	-0.389	0.769
μ_{ND}	9.06	0.833	-0.389	4.604	-0.431
σ_{NS}	23.04	-0.417	0.769	-0.431	1.513

Table 2.6: Estimates and bootstrap standard errors and correlation matrices for the S&P 100 and NASDAQ 100. The top panel uses daily return data between January 3, 1983, and December 31, 2007 (6,307 days) to estimate the parameter values in the left-most column. The rightmost 4 columns contain the bootstrap standard errors (diagonal elements) and the correlations (off-diagonal elements). The bottom panel contains estimates, bootstrap standard errors and correlations from monthly data between January 1983 and December 2007 (300 months). All parameter and covariance estimates have been annualized.

where

$$\mathbf{g}_n(\tilde{\theta}) = n^{-1} \sum_{i=1}^n \mathbf{g}_i(\tilde{\theta})$$

and where $\tilde{\mu}_{SP} = \hat{\mu}_{SP}$, $\tilde{\mu}_{ND} = \hat{\mu}_{ND}$ (unchanged) and $\tilde{\sigma}_{SP}^2 = \tilde{\sigma}_{ND}^2 = (\hat{\sigma}_{SP}^2 + \hat{\sigma}_{ND}^2) / 2$.

2.7.2.1 Bootstrap Covariance Estimation

The bootstrap is an alternative to the plug-in covariance estimators. The bootstrap was implemented using 10,000 resamples where the data were assumed to be i.i.d.. In each bootstrap resample, the full 4 by 1 vector of parameters was computed. These were combined to estimate the parameter covariance using

$$\hat{\Sigma} = B^{-1} \sum_{i=1}^B (\tilde{\theta}_b - \hat{\theta}) (\tilde{\theta}_b - \hat{\theta})'$$

Table 2.6 contains the bootstrap standard errors and correlations. Like the results in 2.5, the parameter estimates and covariance have been annualized, and volatility rather than variance is reported. The covariance estimates are virtually indistinguishable to those computed using the plug-in estimator.

This highlights that the bootstrap is not (generally) a better estimator, but is merely an alternative.²⁸

2.7.3 Testing Factor Exposure

Suppose excess returns were conditionally normal with mean $\mu_i = \beta' \mathbf{x}_i$ and constant variance σ^2 . This type of model is commonly used to explain cross-sectional variation in returns, and when the conditioning variables include only the market variable, the model is known as the Capital Asset Pricing Model (CAP-M, Sharpe (1964) and Lintner (1965)). Multi-factor models allow for additional conditioning variables such as the size and value factors (Ross, 1976; Fama and French, 1992; Fama and French, 1993). The size factor is the return on a portfolio which is long small cap stocks and short large cap stocks. The value factor is the return on a portfolio that is long high book-to-market stocks (value) and short low book-to-market stocks (growth).

This example estimates a 3 factor model where the conditional mean of excess returns on individual assets is modeled as a linear function of the excess return to the market, the size factor and the value factor. This leads to a model of the form

$$\begin{aligned} r_i - r_i^f &= \beta_0 + \beta_1 (r_{m,i} - r_i^f) + \beta_2 r_{s,i} + \beta_3 r_{v,i} + \varepsilon_i \\ r_i^e &= \beta' \mathbf{x}_i + \varepsilon_i \end{aligned}$$

where r_i^f is the risk-free rate (short term government rate), $r_{m,i}$ is the return to the market portfolio, $r_{s,i}$ is the return to the size portfolio and $r_{v,i}$ is the return to the value portfolio. ε_i is a residual which is assumed to have a $N(0, \sigma^2)$ distribution.

Factor models can be formulated as a conditional maximum likelihood problem,

$$l(\mathbf{r} | \mathbf{X}; \theta) = -\frac{1}{2} \sum_{i=1}^n \left\{ \ln(2\pi) + \ln(\sigma^2) + \frac{(r_i - \beta' \mathbf{x}_i)^2}{\sigma^2} \right\}$$

where $\theta = [\beta' \sigma^2]'$. The MLE can be found using the first order conditions, which are

$$\begin{aligned} \frac{\partial l(r; \theta)}{\partial \beta} &= \frac{1}{\hat{\sigma}^2} \sum_{i=1}^n \mathbf{x}_i (r_i - \hat{\beta}' \mathbf{x}_i) = 0 \\ \Rightarrow \hat{\beta} &= \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{j=1}^n \mathbf{x}_i r_i \\ \frac{\partial l(r; \theta)}{\partial \sigma^2} &= -\frac{1}{2} \sum_{i=1}^n \frac{1}{\hat{\sigma}^2} - \frac{(r_i - \hat{\beta}' \mathbf{x}_i)^2}{\hat{\sigma}^4} = 0 \\ \Rightarrow \hat{\sigma}^2 &= n^{-1} \sum_{i=1}^n (r_i - \hat{\beta}' \mathbf{x}_i)^2 \end{aligned}$$

²⁸In this particular application, as the bootstrap and the plug-in estimators are identical as $B \rightarrow \infty$ for fixed n . This is not generally the case.

The vector of scores is

$$\frac{\partial l(r_i|\mathbf{x}_i; \theta)}{\partial \theta} = \begin{bmatrix} \frac{1}{\sigma^2} \mathbf{x}_i \varepsilon_i \\ -\frac{1}{2\sigma^2} + \frac{\varepsilon_i^2}{2\sigma^4} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix} \begin{bmatrix} \mathbf{x}_i \varepsilon_i \\ \sigma^2 - \varepsilon_i^2 \end{bmatrix} = \mathbf{S} \begin{bmatrix} \mathbf{x}_i \varepsilon_i \\ \sigma^2 - \varepsilon_i^2 \end{bmatrix}$$

where $\varepsilon_i = r_i - \beta' \mathbf{x}_i$. The second form will be used to simplify estimating the parameters covariance. The Hessian is

$$\frac{\partial^2 l(r_i|\mathbf{x}_i; \theta)}{\partial \theta \partial \theta'} = \begin{bmatrix} -\frac{1}{\sigma^2} \mathbf{x}_i \mathbf{x}'_i & -\frac{1}{\sigma^4} \mathbf{x}_i \varepsilon_i \\ -\frac{1}{\sigma^4} \mathbf{x}_i \varepsilon_i & \frac{1}{2\sigma^4} - \frac{\varepsilon_i^2}{\sigma^6} \end{bmatrix},$$

and the information matrix is

$$\begin{aligned} \mathcal{I} &= -E \begin{bmatrix} -\frac{1}{\sigma^2} \mathbf{x}_i \mathbf{x}'_i & -\frac{1}{\sigma^4} \mathbf{x}_i \varepsilon_i \\ -\frac{1}{\sigma^4} \mathbf{x}_i \varepsilon_i & \frac{1}{2\sigma^4} - \frac{\varepsilon_i^2}{\sigma^6} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{\sigma^2} E[\mathbf{x}_i \mathbf{x}'_i] & -\frac{1}{\sigma^4} E[\mathbf{x}_i E[\varepsilon_i | \mathbf{X}]] \\ -\frac{1}{\sigma^4} E[\mathbf{x}_i E[\varepsilon_i | \mathbf{X}]] & E\left[\frac{1}{2\sigma^4}\right] \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{\sigma^2} E[\mathbf{x}_i \mathbf{x}'_i] & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}. \end{aligned}$$

The covariance of the scores is

$$\begin{aligned} \mathcal{J} &= E \begin{bmatrix} \varepsilon_i^2 \mathbf{x}_i \mathbf{x}'_i & \sigma^2 \mathbf{x}_i \varepsilon_i - \mathbf{x}_i \varepsilon_i^3 \\ \sigma^2 \mathbf{x}'_i \varepsilon_i - \mathbf{x}'_i \varepsilon_i^3 & (\sigma^2 - \varepsilon_i^2)^2 \end{bmatrix} \mathbf{S} \\ &= \mathbf{S} \begin{bmatrix} E[\varepsilon_i^2 \mathbf{x}_i \mathbf{x}'_i] & E[\sigma^2 \mathbf{x}_i \varepsilon_i - \mathbf{x}_i \varepsilon_i^3] \\ E[\sigma^2 \mathbf{x}'_i \varepsilon_i - \mathbf{x}'_i \varepsilon_i^3] & E[(\sigma^2 - \varepsilon_i^2)^2] \end{bmatrix} \mathbf{S} \\ &= \mathbf{S} \begin{bmatrix} E[E[\varepsilon_i^2 | \mathbf{X}] \mathbf{x}_i \mathbf{x}'_i] & E[\sigma^2 \mathbf{x}'_i E[\varepsilon_i | \mathbf{X}] - \mathbf{x}'_i E[\varepsilon_i^3 | \mathbf{X}]] \\ E[E[\sigma^2 \mathbf{x}'_i \varepsilon_i - \mathbf{x}'_i \varepsilon_i^3 | \mathbf{X}]] & E[(\sigma^2 - \varepsilon_i^2)^2] \end{bmatrix} \mathbf{S} \\ &= \mathbf{S} \begin{bmatrix} \sigma^2 E[\mathbf{x}_i \mathbf{x}'_i] & 0 \\ 0 & 2\sigma^4 \end{bmatrix} \mathbf{S} = \begin{bmatrix} \frac{1}{\sigma^2} E[\mathbf{x}_i \mathbf{x}'_i] & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix} \end{aligned}$$

The estimators of the covariance matrices are

$$\begin{aligned} \hat{\mathcal{J}} &= n^{-1} \sum_{i=1}^n \begin{bmatrix} \frac{1}{\hat{\sigma}^2} & 0 \\ 0 & \frac{1}{2\hat{\sigma}^4} \end{bmatrix} \begin{bmatrix} \mathbf{x}_i \hat{\varepsilon}_i \\ \hat{\sigma}^2 - \hat{\varepsilon}_i^2 \end{bmatrix} \begin{bmatrix} \mathbf{x}'_i \hat{\varepsilon}_i & \hat{\sigma}^2 - \hat{\varepsilon}_i^2 \end{bmatrix} \begin{bmatrix} \frac{1}{\hat{\sigma}^2} & 0 \\ 0 & \frac{1}{2\hat{\sigma}^4} \end{bmatrix} \\ &= n^{-1} \sum_{i=1}^n \begin{bmatrix} \frac{1}{\hat{\sigma}^2} & 0 \\ 0 & \frac{1}{2\hat{\sigma}^4} \end{bmatrix} \begin{bmatrix} \hat{\varepsilon}_i^2 \mathbf{x}_i \mathbf{x}'_i & \hat{\sigma}^2 \mathbf{x}_i \hat{\varepsilon}_i - \mathbf{x}_i \hat{\varepsilon}_i^3 \\ \hat{\sigma}^2 \mathbf{x}'_i \hat{\varepsilon}_i - \mathbf{x}'_i \hat{\varepsilon}_i^3 & (\hat{\sigma}^2 - \hat{\varepsilon}_i^2)^2 \end{bmatrix} \begin{bmatrix} \frac{1}{\hat{\sigma}^2} & 0 \\ 0 & \frac{1}{2\hat{\sigma}^4} \end{bmatrix} \end{aligned}$$

and

$$\begin{aligned} \hat{\mathcal{I}} &= -1 \times n^{-1} \sum_{i=1}^n \begin{bmatrix} -\frac{1}{\hat{\sigma}^2} \mathbf{x}_i \mathbf{x}'_i & -\frac{1}{\hat{\sigma}^4} \mathbf{x}_i \varepsilon_i \\ -\frac{1}{\hat{\sigma}^4} \mathbf{x}_i \varepsilon_i & \frac{1}{2\hat{\sigma}^4} - \frac{\varepsilon_i^2}{\hat{\sigma}^6} \end{bmatrix} \\ &= -1 \times n^{-1} \sum_{i=1}^n \begin{bmatrix} -\frac{1}{\hat{\sigma}^2} \mathbf{x}_i \mathbf{x}'_i & 0 \\ 0 & \frac{1}{2\hat{\sigma}^4} - \frac{\hat{\sigma}^2}{\hat{\sigma}^6} \end{bmatrix} \\ &= -1 \times n^{-1} \sum_{i=1}^n \begin{bmatrix} -\frac{1}{\hat{\sigma}^2} \mathbf{x}_i \mathbf{x}'_i & 0 \\ 0 & -\frac{1}{2\hat{\sigma}^4} \end{bmatrix} = n^{-1} \sum_{i=1}^n \begin{bmatrix} \frac{1}{\hat{\sigma}^2} & 0 \\ 0 & \frac{1}{2\hat{\sigma}^4} \end{bmatrix} \begin{bmatrix} \mathbf{x}_i \mathbf{x}'_i & 0 \\ 0 & 1 \end{bmatrix} \end{aligned}$$

Note that the off-diagonal term in \mathcal{J} , $\hat{\sigma}^2 \mathbf{x}_i' \hat{\varepsilon}_i - \mathbf{x}_i' \hat{\varepsilon}_i^3$, is not necessarily 0 when the data may be conditionally skewed. Combined, the QMLE parameter covariance estimator is then

$$\begin{aligned}\hat{\mathcal{I}}^{-1} \hat{\mathcal{J}} \hat{\mathcal{I}}^{-1} &= \left(n^{-1} \sum_{i=1}^n \begin{bmatrix} \mathbf{x}_i \mathbf{x}_i' & 0 \\ 0 & 1 \end{bmatrix} \right)^{-1} \left[n^{-1} \sum_{i=1}^n \begin{bmatrix} \hat{\varepsilon}_i^2 \mathbf{x}_i \mathbf{x}_i' & \hat{\sigma}^2 \mathbf{x}_i \hat{\varepsilon}_i - \mathbf{x}_i' \hat{\varepsilon}_i^3 \\ \hat{\sigma}^2 \mathbf{x}_i' \hat{\varepsilon}_i - \mathbf{x}_i' \hat{\varepsilon}_i^3 & (\hat{\sigma}^2 - \hat{\varepsilon}_i^2)^2 \end{bmatrix} \right] \\ &\quad \times \left(n^{-1} \sum_{i=1}^n \begin{bmatrix} \mathbf{x}_i \mathbf{x}_i' & 0 \\ 0 & 1 \end{bmatrix} \right)^{-1}\end{aligned}$$

where the identical scaling terms have been canceled. Additionally, when returns are conditionally normal,

$$\begin{aligned}\text{plim } \hat{J} &= \text{plim } n^{-1} \sum_{i=1}^n \begin{bmatrix} \frac{1}{\hat{\sigma}^2} & 0 \\ 0 & \frac{1}{2\hat{\sigma}^4} \end{bmatrix} \begin{bmatrix} \hat{\varepsilon}_i^2 \mathbf{x}_i \mathbf{x}_i' & \hat{\sigma}^2 \mathbf{x}_i \hat{\varepsilon}_i - \mathbf{x}_i' \hat{\varepsilon}_i^3 \\ \hat{\sigma}^2 \mathbf{x}_i' \hat{\varepsilon}_i - \mathbf{x}_i' \hat{\varepsilon}_i^3 & (\hat{\sigma}^2 - \hat{\varepsilon}_i^2)^2 \end{bmatrix} \begin{bmatrix} \frac{1}{\hat{\sigma}^2} & 0 \\ 0 & \frac{1}{2\hat{\sigma}^4} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{\hat{\sigma}^2} & 0 \\ 0 & \frac{1}{2\hat{\sigma}^4} \end{bmatrix} \begin{bmatrix} \sigma^2 \mathbf{x}_i \mathbf{x}_i' & 0 \\ 0 & 2\sigma^4 \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{\sigma^2} \mathbf{x}_i \mathbf{x}_i' & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix}\end{aligned}$$

and

$$\begin{aligned}\text{plim } \hat{\mathcal{I}} &= \text{plim } n^{-1} \sum_{i=1}^n \begin{bmatrix} \frac{1}{\hat{\sigma}^2} \mathbf{x}_i \mathbf{x}_i' & 0 \\ 0 & \frac{1}{2\hat{\sigma}^4} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{\sigma^2} \mathbf{x}_i \mathbf{x}_i' & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix},\end{aligned}$$

and so the IME, $\text{plim } \hat{\mathcal{J}} - \hat{\mathcal{I}} = 0$, will hold when returns are conditionally normal. Moreover, when returns are not normal, all of the terms in \mathcal{J} will typically differ from the limits above and so the IME will not generally hold.

2.7.3.1 Data and Implementation

Three assets are used to illustrate hypothesis testing: ExxonMobil (XOM), Google (GOOG) and the SPDR Gold Trust ETF (GLD). The data used to construct the individual equity returns were downloaded from Yahoo! Finance and span the period September 2, 2002, until September 1, 2012.²⁹ The market portfolio is the CRSP value-weighted market, which is a composite based on all listed US equities. The size and value factors were constructed using portfolio sorts and are made available by Ken French. All returns were scaled by 100.

2.7.3.2 Wald tests

Wald tests make use of the parameters and estimated covariance to assess the evidence against the null. When testing whether the size and value factor are relevant for an asset, the null is $H_0 : \beta_2 = \beta_3 = 0$.

²⁹Google and the SPDR Gold Trust ETF both started trading after the initial sample date. In both cases, all available data was used.

This problem can be set up as a Wald test using

$$\mathbf{R} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \mathbf{r} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

and

$$W = n(\mathbf{R}\hat{\theta} - \mathbf{r})' [\mathbf{R}\hat{\mathcal{I}}^{-1}\mathcal{J}\hat{\mathcal{I}}^{-1}\mathbf{R}']^{-1} (\mathbf{R}\hat{\theta} - \mathbf{r}).$$

The Wald test has an asymptotic χ^2_2 distribution since the null imposes 2 restrictions.

t-stats can similarly be computed for individual parameters

$$t_j = \sqrt{n} \frac{\hat{\beta}_j}{\text{s.e.}(\hat{\beta}_j)}$$

where s.e. $(\hat{\beta}_j)$ is the square of the j^{th} diagonal element of the parameter covariance matrix. Table 2.7 contains the parameter estimates from the models, *t*-stats for the coefficients and the Wald test statistics for the null $H_0 : \beta_2 = \beta_3 = 0$. The *t*-stats and the Wald tests were implemented using both the sandwich covariance estimator (QMLE) and the maximum likelihood covariance estimator. The two sets of test statistics differ in magnitude since the assumption of normality is violated in the data, and so only the QMLE-based test statistics should be considered reliable.

2.7.3.3 Likelihood Ratio tests

Likelihood ratio tests are simple to implement when parameters are estimated using MLE. The likelihood ratio test statistic is

$$LR = -2(l(\mathbf{r}|\mathbf{X}; \tilde{\theta}) - l(\mathbf{r}|\mathbf{X}; \hat{\theta}))$$

where $\tilde{\theta}$ is the null-restricted estimator of the parameters. The likelihood ratio has an asymptotic χ^2_2 distribution since there are two restrictions. Table 2.7 contains the likelihood ratio test statistics for the null $H_0 : \beta_2 = \beta_3 = 0$. Caution is needed when interpreting likelihood ratio test statistics since the asymptotic distribution is only valid when the model is correctly specified – in this case, when returns are conditionally normal, which is not plausible.

2.7.3.4 Lagrange Multiplier tests

Lagrange Multiplier tests are somewhat more involved in this problem. The key to computing the LM test statistic is to estimate the score using the restricted parameters,

$$\tilde{\mathbf{s}}_i = \begin{bmatrix} \frac{1}{\sigma^2} \mathbf{x}_i \tilde{\varepsilon}_i \\ -\frac{1}{2\sigma^2} + \frac{\tilde{\varepsilon}_i^2}{2\sigma^4} \end{bmatrix},$$

where $\tilde{\varepsilon}_i = r_i - \tilde{\beta}' \mathbf{x}_i$ and $\tilde{\theta} = [\tilde{\beta}' \tilde{\sigma}^2]'$ is the vector of parameters estimated when the null is imposed. The LM test statistic is then

$$LM = n\bar{\mathbf{s}}\tilde{\mathbf{S}}^{-1}\bar{\mathbf{s}}$$

where

$$\bar{\mathbf{s}} = n^{-1} \sum_{i=1}^n \tilde{\mathbf{s}}_i, \text{ and } \tilde{\mathbf{S}} = n^{-1} \sum_{i=1}^n \tilde{\mathbf{s}}_i \tilde{\mathbf{s}}_i'.$$

The improved version of the LM can be computed by replacing $\tilde{\mathbf{S}}$ with a covariance estimator based on the scores from the unrestricted estimates,

$$\hat{\mathbf{S}} = n^{-1} \sum_{i=1}^n \hat{\mathbf{s}}_i \hat{\mathbf{s}}_i'.$$

Table 2.7 contains the LM test statistics for the null $H_0 : \beta_2 = \beta_3 = 0$ using the two covariance estimators. LM test statistics are naturally robust to violations of the assumed normality since $\hat{\mathbf{S}}$ and $\tilde{\mathbf{S}}$ are directly estimated from the scores and not based on properties of the assumed normal distribution.

2.7.3.5 Discussion of Test Statistics

Table 2.7 contains all test statistics for the three series. The test statistics based on the MLE and QMLE parameter covariances differ substantially in all three series, and importantly, the conclusions also differ for the SPDR Gold Trust ETF. The difference between the two sets of results from an implicit rejection of the assumption that returns are conditionally normal with constant variance. The MLE-based Wald test and the LR test (which is implicitly MLE-based) have very similar magnitudes for all three series. The QMLE-based Wald test statistics are also always larger than the LM-based test statistics which reflects the difference of estimating the covariance under the null or under the alternative.

Shorter Problems

Problem 2.1. What influences the power of a hypothesis test?

Problem 2.2. Let Y_i be i.i.d. $\text{Exponential}(\lambda)$ with pdf $f(y_i) = \lambda \exp(-\lambda y_i)$, $\lambda > 0$. Derive the MLE of λ where there are n observations.

Problem 2.3. If n observations of $Y_i \stackrel{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$ are observed, what is the MLE of p ? The pdf of a single Bernoulli is

$$p^y (1-p)^{1-y}.$$

Problem 2.4. When performing a hypothesis test, what are Type I and Type II Errors?

Longer Exercises

Exercise 2.1. The distribution of a discrete random variable X depends on a discretely valued parameter $\theta \in \{1, 2, 3\}$ according to

x	$f(x \theta = 1)$	$f(x \theta = 2)$	$f(x \theta = 3)$
1	$\frac{1}{2}$	$\frac{1}{3}$	0
2	$\frac{1}{3}$	$\frac{1}{4}$	0
3	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{6}$
4	0	$\frac{1}{12}$	$\frac{1}{12}$
5	0	0	$\frac{3}{4}$

Find the MLE of θ if one value from X has been observed. Note: The MLE is a function that returns

ExxonMobil					
Parameter	Estimate	t (MLE)	t (QMLE)		
β_0	0.016	0.774 (0.439)	0.774 (0.439)	Wald (MLE)	251.21 (<0.001)
β_1	0.991	60.36 (<0.001)	33.07 (<0.001)		88.00 (<0.001)
β_2	-0.536	-15.13 (<0.001)	-9.24 (<0.001)		239.82 (<0.001)
β_3	-0.231	-6.09 (<0.001)	-3.90 (<0.001)		LM ($\tilde{\mathbf{S}}$) 53.49 LM ($\hat{\mathbf{S}}$) 54.63 (<0.001)
Google					
Parameter	Estimate	t (MLE)	t (QMLE)		
β_0	0.063	1.59 (0.112)	1.60 (0.111)	Wald (MLE)	18.80 (<0.001)
β_1	0.960	30.06 (<0.001)	23.74 (<0.001)		10.34 (0.006)
β_2	-0.034	-0.489 (0.625)	-0.433 (0.665)		18.75 (<0.001)
β_3	-0.312	-4.34 (<0.001)	-3.21 (0.001)		LM ($\tilde{\mathbf{S}}$) 10.27 LM ($\hat{\mathbf{S}}$) 10.32 (0.006)
SPDR Gold Trust ETF					
Parameter	Estimate	t (MLE)	t (QMLE)		
β_0	0.057	1.93 (0.054)	1.93 (0.054)	Wald (MLE)	12.76 (0.002)
β_1	0.130	5.46 (<0.001)	2.84 (0.004)		5.16 (0.076)
β_2	-0.037	-0.733 (0.464)	-0.407 (0.684)		12.74 (0.002)
β_3	-0.191	-3.56 (<0.001)	-2.26 (0.024)		LM ($\tilde{\mathbf{S}}$) 5.07 LM ($\hat{\mathbf{S}}$) 5.08 (0.079)

Table 2.7: Parameter estimates, t-statistics (both MLE and QMLE-based), and tests of the exclusion restriction that the size and value factors have no effect ($H_0 : \beta_2 = \beta_3 = 0$) on the returns of the ExxonMobil, Google and SPDR Gold Trust ETF.

an estimate of θ given the data that has been observed. In the case where both the observed data and the parameter are discrete, a “function” will take the form of a table.

Exercise 2.2. Let X_1, \dots, X_n be an i.i.d. sample from a $\text{gamma}(\alpha, \beta)$ distribution. The density of a $\text{gamma}(\alpha, \beta)$ is

$$f(x; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp(-x/\beta)$$

where $\Gamma(z)$ is the gamma-function evaluated at z . Find the MLE of β assuming α is known.

Exercise 2.3. Let X_1, \dots, X_n be an i.i.d. sample from the pdf

$$f(x|\theta) = \frac{\theta}{x^{\theta+1}}, \quad 1 \leq x < \infty, \theta > 1$$

1. What is the MLE of θ ?
2. What is $E[X_j]$?
3. How can the previous answer be used to compute a method of moments estimator of θ ?

Exercise 2.4. Let X_1, \dots, X_n be an i.i.d. sample from the pdf

$$f(x|\theta) = \frac{1}{\theta}, \quad 0 \leq x \leq \theta, \theta > 0$$

1. What is the MLE of θ ? [This is tricky]
2. What is the method of moments Estimator of θ ?
3. Compute the bias and variance of each estimator.

Exercise 2.5. Let X_1, \dots, X_n be an i.i.d. random sample from the pdf

$$f(x|\theta) = \theta x^{\theta-1}, \quad 0 \leq x \leq 1, 0 < \theta < \infty$$

1. What is the MLE of θ ?
2. What is the variance of the MLE?
3. Show that the MLE is consistent.

Exercise 2.6. Let X_1, \dots, X_i be an i.i.d. sample from a $\text{Bernoulli}(p)$.

1. Show that \bar{X} achieves the Cramér-Rao lower bound.
2. What do you conclude about using \bar{X} to estimate p ?

Exercise 2.7. Suppose you witness a coin being flipped 100 times with 56 heads and 44 tails. Is there evidence that this coin is unfair?

Exercise 2.8. Let X_1, \dots, X_i be an i.i.d. sample with mean μ and variance σ^2 .

1. Show $\tilde{X} = \sum_{i=1}^N w_i X_i$ is unbiased if and only if $\sum_{i=1}^N w_i = 1$.
2. Show that the variance of \tilde{X} is minimized if $w_i = \frac{1}{n}$ for $i = 1, 2, \dots, n$.

Exercise 2.9. Suppose $\{X_i\}$ is i.i.d. sequence of normal variables with unknown mean μ and known variance σ^2 .

1. Derive the power function of a 2-sided t -test of the null $H_0 : \mu = 0$ against an alternative $H_1 : \mu \neq 0$? The power function should have two arguments, the mean under the alternative, μ_1 and the number of observations n .
2. Sketch the power function for $n = 1, 4, 16, 64, 100$.
3. What does this tell you about the power as $n \rightarrow \infty$ for $\mu \neq 0$?

Exercise 2.10. Let X_1 and X_2 are independent and drawn from a $\text{Uniform}(\theta, \theta + 1)$ distribution with θ unknown. Consider two test statistics,

$$T_1 : \text{Reject if } X_1 > .95$$

and

$$T_2 : \text{Reject if } X_1 + X_2 > C$$

1. What is the size of T_1 ?
2. What value must C take so that the size of T_2 is equal to T_1
3. Sketch the power curves of the two tests as a function of θ . Which is more powerful?

Exercise 2.11. Suppose $\{y_i\}$ are a set of transaction counts (trade counts) over 5-minute intervals which are believed to be i.i.d. distributed from a Poisson with parameter λ . Recall the probability density function of a Poisson is

$$f(y_i; \lambda) = \frac{\lambda^{y_i} e^{-\lambda}}{y_i!}$$

1. What is the log-likelihood for this problem?
2. What is the MLE of λ ?
3. What is the variance of the MLE?
4. Suppose that $\hat{\lambda} = 202.4$ and that the sample size was 200. Construct a 95% confidence interval for λ .
5. Use a t -test to test the null $H_0 : \lambda = 200$ against $H_1 : \lambda \neq 200$ with a size of 5%
6. Use a likelihood ratio to test the same null with a size of 5%.
7. What happens if the assumption of i.i.d. data is correct but that the data does not follow a Poisson distribution?

**Upper tail probabilities
for a standard normal z**

Cut-off c	$\Pr(z > c)$
1.282	10%
1.645	5%
1.96	2.5%
2.32	1%

5% Upper tail cut-off for χ_q^2

Degree of Freedom q	Cut-Off
1	3.84
2	5.99
199	232.9
200	234.0

Exercise 2.12. Suppose $Y_i|X_i = x_i \stackrel{\text{i.i.d.}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$

1. Write down the log-likelihood for this problem.
 2. Find the MLE of the unknown parameters.
 3. What is the asymptotic distribution of the parameters?
 4. Describe two classes tests to test the null $H_0 : \beta_1 = 0$ against the alternative $H_0 : \beta_1 \neq 0$.
 5. How would you test whether the errors in the model were conditionally heteroskedastic?
 6. Suppose $X_i \stackrel{\text{i.i.d.}}{\sim} N(\mu_X, \sigma_X^2)$ and the X variables are independent of the shocks in the model. What are the values of:
 - (a) $E[Y_i]$
 - (b) $E[Y_i^2]$
 - (c) $V[Y_i]$
 - (d) $\text{Cov}[X_i, Y_i]$
- Note: If $Y \sim N(\mu, \sigma^2)$, then the pdf of Y is

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

Exercise 2.13. Suppose $Y_i \stackrel{\text{i.i.d.}}{\sim} \text{Exponential}(\lambda)$, so that $E[Y_i] = \lambda$.

1. Write down the log-likelihood for this problem.
2. Find the MLE of the unknown parameter.
3. What is the asymptotic distribution of the parameter estimate?

4. Suppose $n = 10$, $\sum y_i = 19$. Test the null $H_0 : \lambda = 1$ against a 2-sided alternative with a size of 5% test using a t-test.
5. Suppose $n = 10$, $\sum y_i = 19$. Test the null $H_0 : \lambda = 1$ against a 2-sided alternative with a size of 5% test using a likelihood-ratio.
6. When are sandwich covariance estimators needed in MLE problems?
7. Discuss the important considerations for building models using cross-sectional data?

Notes:

- If $Y \sim \text{Exponential}(\lambda)$, then the pdf of Y is

$$f(y; \lambda) = \frac{1}{\lambda} \exp\left(-\frac{y}{\lambda}\right)$$

- The 5% critical value for a χ^2_1 is 3.8415, for a χ^2_2 is 5.9915 and for a χ^2_3 is 7.8147.

Exercise 2.14. Suppose $y_i|x_i \sim \text{Exponential}(x_i\beta)$ where $x_i > 0$ and $\beta > 0$. This can be equivalently written $y_i \sim \text{Exponential}(\lambda_i)$ where $\lambda_i = x_i\beta$. The PDF of an exponential random variance with parameter λ is

$$f_Y(y) = \lambda \exp(-\lambda y).$$

Assume n pairs of observations on (y_i, x_i) are observed

1. What is the log-likelihood of the data?
2. Compute the maximum likelihood estimator $\hat{\beta}$.
3. What is the asymptotic distribution of $\sqrt{n}(\hat{\beta} - \beta)$?
4. Suppose the following quantities are observed

$$\begin{aligned} n &= 20 \\ \sum_{i=1}^n x_i &= 16.58 \\ \sum_{i=1}^n y_i &= 128.47 \\ \sum_{i=1}^n x_i y_i &= 11.23 \end{aligned}$$

Perform a test for the null $H_0 : \beta = 1.5$ against the alternative $H_1 : \beta \neq 1.5$ using a t-test.

5. Explain how you would perform a likelihood-ratio test for the same null and alternative.

Chapter 3

Analysis of Cross-Sectional Data

Note: The primary reference text for these notes is Hayashi (2000). Other comprehensive treatments are available in Greene (2007) and Davidson and MacKinnon (2003).

Linear regression is the foundation of modern econometrics. While the importance of linear regression in financial econometrics has diminished in recent years, it is still widely employed. More importantly, the theory behind least-squares estimators is useful in broader contexts, and many results of this chapter are special cases of more general estimators presented in subsequent chapters. This chapter covers model specification, estimation, small- and large-sample inference, and model selection.

Linear regression is an essential tool of any econometrician and is widely used throughout finance and economics. Linear regression's success is owed to two key features: the availability of simple, closed-form estimators, and the ease and directness of interpretation. However, despite the regression estimator's superficial simplicity, the concepts presented in this chapter will reappear in the chapters on time series, panel data, Generalized Method of Moments (GMM), event studies, and volatility modeling.

3.1 Model Description

Linear regression expresses a dependent variable as a linear function of independent variables, possibly random, and an error.

$$Y_i = \beta_1 X_{1,i} + \beta_2 X_{2,i} + \dots + \beta_k X_{k,i} + \varepsilon_i, \quad (3.1)$$

where Y_i is known as the *regressand*, *dependent variable* or simply the *left-hand-side variable*. The k variables, $X_{1,i}, \dots, X_{k,i}$ are known as the *regressors*, *independent variables* or *right-hand-side variables*. $\beta_1, \beta_2, \dots, \beta_k$ are the *regression coefficients*, ε_i is known as the *innovation*, *shock* or *error* and $i = 1, 2, \dots, n$ index the observation. While this representation clarifies the relationship between Y_i and the X s, matrix notation will generally be used to compactly describe models:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1k} \\ X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad (3.2)$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (3.3)$$

where \mathbf{X} is an n by k matrix, $\boldsymbol{\beta}$ is a k by 1 vector, and both \mathbf{y} and $\boldsymbol{\varepsilon}$ are n by 1 vectors.

Two vector notations will occasionally be used: row,

$$\begin{bmatrix} Y_1 & = & \mathbf{X}_1\boldsymbol{\beta} & + \varepsilon_1 \\ Y_2 & = & \mathbf{X}_2\boldsymbol{\beta} & + \varepsilon_2 \\ \vdots & & \vdots & \vdots \\ Y_n & = & \mathbf{X}_n\boldsymbol{\beta} & + \varepsilon_n \end{bmatrix} \quad (3.4)$$

and column,

$$\mathbf{y} = \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \dots + \beta_k \mathbf{x}_k + \boldsymbol{\varepsilon}. \quad (3.5)$$

Linear regression allows coefficients to be interpreted, *all things being equal*. Specifically, the effect of a change in one variable can be examined without changing the others. Regression analysis also allows for models that contain all of the information relevant for determining Y_i , whether these quantities are of primary interest or not. This feature provides the mechanism to interpret the coefficient on a regressor as the unique effect of that regressor (under certain conditions), a feature that makes linear regression very attractive.

3.1.1 What is a model?

What constitutes a model is a difficult question to answer. One view of a model is that of the *data generating process* (DGP). For instance, if a model postulates

$$Y_i = \beta_1 X_i + \varepsilon_i$$

then one interpretation is that the regressand, Y_i , is wholly determined by X_i and some random shock. The alternative view is that X_i is the only relevant variable available to the econometrician that explains variation in Y_i . Everything else that determines Y_i cannot be measured and, in the usual case, cannot be placed into a framework that would allow the researcher to formulate a model.

Consider monthly returns on the S&P 500, a value-weighted index of 500 large firms in the United States. Equity holdings and returns are generated by individuals based on their beliefs and preferences. If one were to take a (literal) data generating process view of the return on this index, data on individual investors' preferences and beliefs would need to be collected and formulated into a model for market returns. Collecting data and building this model would be a substantial challenge.

On the other hand, a model can be built to explain the variation in the market based on observable quantities (such as oil price changes or macroeconomic news announcements) without explicitly collecting information on beliefs and preferences. In a model of this type, explanatory variables can be viewed as inputs individuals consider when forming their beliefs and, subject to their preferences,

taking actions that ultimately affect the price of the S&P 500. The model allows the relationships between the regressand and regressors to be explored and is meaningful even though the model is not plausibly the data generating process.

In the context of time-series data, models often postulate that a series's past values are useful in predicting future values. Consider building a model of monthly returns on the S&P 500 using past returns to explain future returns. Treated as a DGP, this model implies that average returns in the future are determined by returns in the immediate past. Alternatively, if treated as an approximation, then one interpretation postulates that changes in risk aversion, beliefs, or other variables that influence holdings of assets change slowly (possibly in an unobservable manner). These slowly changing "factors" produce predictability in returns. Of course, there are other interpretations, but these should come from finance theory rather than data. The *model as a proxy* interpretation is additionally useful as it allows models to be specified, which are only loosely coupled with theory but that capture essential features of a theoretical model.

Careful consideration of what defines a model is a crucial step in the development of an econometrician, and one should always consider which assumptions and beliefs are needed to justify any specification.

3.1.2 Example: Cross-section regression of returns on factors

The concepts of linear regression will be explored in the context of a cross-section regression of returns on a set of factors thought to capture systematic risk. Cross-sectional regressions in financial econometrics date back at least to the Capital Asset Pricing Model (CAPM, Markowitz (1959), Sharpe (1964) and Lintner (1965)), a model formulated as a regression of individual asset's excess returns on the excess return of the market. More general specifications with multiple regressors are motivated by the Intertemporal CAPM (ICAPM, Merton (1973)) and Arbitrage Pricing Theory (APT, Ross (1976)).

The basic model postulates that excess returns are linearly related to a set of systematic risk factors. The factors can be returns on other assets, such as the market portfolio, or any other variable related to intertemporal hedging demands, such as interest rates, shocks to inflation, or consumption growth.

$$R_i - R_i^f = \mathbf{f}_i \boldsymbol{\beta} + \varepsilon_i$$

or more compactly,

$$r_i^e = \mathbf{f}_i \boldsymbol{\beta} + \varepsilon_i$$

where $R_i^e = R_i - R_i^f$ is the excess return on the asset and $\mathbf{f}_i = [F_{1,i}, \dots, F_{k,i}]$ are returns on factors that explain systematic variation.

Linear factors models have been used in countless studies, the most well known by Fama and French (Fama and French (1992) and Fama and French (1993)) who use returns on specially constructed portfolios as factors to capture specific types of risk. The data set contains the variables listed in table 3.1.

Monthly data from July 1963 until January 2020 is used in the examples. Except for the interest rates, all return data are from the CRSP database. Returns are calculated as 100 times the logarithmic price difference ($R_i = 100(\ln(P_i) - \ln(P_{i-1}))$). Portfolios were constructed by sorting the firms into categories based on market capitalization, Book Equity to Market Equity (BE/ME), or past returns

Variable	Description
<i>VWM</i>	Returns on a value-weighted portfolio of all NYSE, AMEX and NASDAQ stocks
<i>SMB</i>	Returns on the Small minus Big factor, a zero investment portfolio that is long small market capitalization firms and short big caps.
<i>HML</i>	Returns on the High minus Low factor, a zero investment portfolio that is long high BE/ME firms and short low BE/ME firms.
<i>MOM</i>	Returns on a portfolio that is long winners and short losers as defined by their performance over the past 12 months, excluding the last month. Includes the large and small cap stocks but excludes mid-cap stocks.
<i>SL</i>	Returns on a portfolio of small cap and low BE/ME firms.
<i>SM</i>	Returns on a portfolio of small cap and medium BE/ME firms.
<i>SH</i>	Returns on a portfolio of small cap and high BE/ME firms.
<i>BL</i>	Returns on a portfolio of big cap and low BE/ME firms.
<i>BM</i>	Returns on a portfolio of big cap and medium BE/ME firms.
<i>BH</i>	Returns on a portfolio of big cap and high BE/ME firms.
<i>RF</i>	Risk free rate (Rate on a 3 month T-bill).
<i>DATE</i>	Date in format YYYYMM.

Table 3.1: Variable description for the data available in the Fama-French data-set used throughout this chapter.

over the previous year. For further details on the construction of portfolios, see Fama and French (1993) or Ken French's website:

http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

A general model for the *BH* portfolio can be specified

$$BH_i - RF_i = \beta_1 + \beta_2(VWM_i - RF_i) + \beta_3SMB_i + \beta_4HML_i + \beta_5MOM_i + \varepsilon_i$$

or, in terms of the excess returns,

$$BH_i^e = \beta_1 + \beta_2VWM_i^e + \beta_3SMB_i + \beta_4HML_i + \beta_5MOM_i + \varepsilon_i.$$

The coefficients in the model can be interpreted as the effect of a change in one variable holding the other variables constant. For example, β_3 captures the effect of a change in the SMB_i risk factor holding VWM_i^e , HML_i and MOM_i constant. Table 3.2 contains some descriptive statistics of the factors and the six portfolios included in this data set.

3.2 Functional Form

A linear relationship is fairly specific and, in some cases, restrictive. It is important to distinguish specifications that can be examined in the linear regression framework from those that cannot. Linear

	Mean	Std. Dev.	Skewness	Kurtosis
<i>VWM^e</i>	6.66	15.42	-0.54	4.91
<i>SMB</i>	2.17	10.52	0.43	7.83
<i>HML</i>	3.06	9.95	0.01	5.41
<i>MOM</i>	7.95	14.52	-1.28	13.20
<i>SL^e</i>	6.54	23.55	-0.39	4.74
<i>SM^e</i>	10.21	18.93	-0.54	5.81
<i>SH^e</i>	11.23	19.69	-0.53	6.80
<i>BL^e</i>	6.78	15.94	-0.34	4.84
<i>BM^e</i>	6.47	14.87	-0.48	5.39
<i>BH^e</i>	8.22	17.20	-0.62	6.23

Table 3.2: Descriptive statistics of the six portfolios that will be used throughout this chapter. The data consist of monthly observations from January 1927 until June 2008 ($n = 978$).

regressions require two key features of any model: each term on the right-hand side must have only one coefficient that enters multiplicatively, and the error must enter additively.¹ Most specifications satisfying these two requirements can be treated using the tools of linear regression.² Other forms of “nonlinearities” are permissible. Any regressor or the regressand can be nonlinear transformations of the original observed data.

Double log (also known as *log-log*) specifications, where both the regressor and the regressands are log transformations of the original (positive) data, are frequently used.

$$\ln Y_i = \beta_1 + \beta_2 \ln X_i + \varepsilon_i.$$

In the parlance of linear regression, the model is specified

$$\tilde{Y}_i = \beta_1 + \beta_2 \tilde{X}_i + \varepsilon_i$$

where $\tilde{Y}_i = \ln(Y_i)$ and $\tilde{X}_i = \ln(X_i)$. The usefulness of the double log specification can be illustrated by a Cobb-Douglas production function subject to a multiplicative shock

$$Y_i = \beta_1 K_i^{\beta_2} L_i^{\beta_3} \varepsilon_i.$$

Using the production function directly, it is not obvious that, given values for output (Y_i), capital (K_i) and labor (L_i) of firm i , the model is consistent with a linear regression. However, taking logs,

$$\ln Y_i = \ln \beta_1 + \beta_2 \ln K_i + \beta_3 \ln L_i + \ln \varepsilon_i$$

the model can be reformulated as a linear regression on the transformed data. Other forms, such as semi-log (either log-lin, where the regressand is logged but the regressors are unchanged, or lin-log, which logs only the regressor), are often useful to describe nonlinear relationships.

¹A third but obvious requirement is that neither Y_i nor any of the $X_{j,i}$ may be latent (unobservable), $j = 1, 2, \dots, k$, $i = 1, 2, \dots, n$.

²There are further requirements on the data, both the regressors and the regressand, to ensure that estimators of the unknown parameters are reasonable, but these are treated in subsequent sections.

Linear regression does, however, rule out specifications that may be of interest. Linear regression is not an appropriate framework to examine a model of the form $Y_i = \beta_1 X_{1,i}^{\beta_2} + \beta_3 X_{2,i}^{\beta_4} + \varepsilon_i$. Fortunately, more general frameworks, such as the generalized method of moments (GMM) or maximum likelihood estimation (MLE), topics of subsequent chapters, can be applied.

Two other transformations of the original data, dummy variables and interactions, are commonly used to generate nonlinear (in regressors) specifications. A *dummy variable* is a special class of regressor that takes the value 0 or 1. In finance, dummy variables (or dummies) are used to model calendar effects, leverage (where the magnitude of a coefficient depends on the sign of the regressor), or group-specific effects. Variable *interactions* parameterize nonlinearities into a model through products of regressors. Common interactions include powers of regressors ($X_{1,i}^2, X_{1,i}^3, \dots$), cross-products of regressors ($X_{1,i}X_{2,i}$) and interactions between regressors and dummy variables. Variable transformations add significant flexibility to the linear regression models.

The use of nonlinear transformations also changes the interpretation of the regression coefficients. If only unmodified regressors are included,

$$Y_i = \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i$$

then $\frac{\partial Y_i}{\partial X_{k,i}} = \beta_k$. Suppose a specification includes both X_i and X_i^2 as regressors,

$$Y_i = \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$

In this specification, $\frac{\partial Y_i}{\partial X_i} = \beta_1 + \beta_2 X_i$ and the level of the variable enters its partial effect. Similarly, in a simple double log model

$$\ln Y_i = \beta_1 \ln X_i + \varepsilon_i,$$

and

$$\beta_1 = \frac{\partial \ln Y_i}{\partial \ln X_i} = \frac{\frac{\partial Y}{\partial X}}{\frac{\partial X}{X}} = \frac{\% \Delta Y}{\% \Delta X}$$

Thus, β_1 corresponds to the *elasticity* of Y_i with respect to X_i . In general, the coefficient on a variable that enters the model in levels corresponds to the effect of a one-unit change in that variable. The coefficient on a variable that appears logged corresponds to the effect of a one percent change in that variable. For example, in a semi-log model where only the regressor is logged,

$$Y_i = \beta_1 \ln X_i + \varepsilon_i,$$

β_1 will correspond to a unit change in Y_i for a % change in X_i . Finally, in the case of discrete regressors, where there is no differential interpretation of coefficients, β represents the effect of a *whole* unit change, such as a dummy going from 0 to 1.

3.2.1 Example: Dummy variables and interactions in cross-section regressions

The January and the December effects are seasonal phenomena that have been widely studied in finance. Simply put, the December effect hypothesizes that returns in December are unusually low

due to tax-induced portfolio rebalancing, mostly to realized losses, while the January effect stipulates returns are abnormally high as investors return to the market. To model excess returns on a portfolio (BH_i^e) as a function of the excess market return (VWM_i^e), a constant, and the January and December effects, a model can be specified

$$BH_i^e = \beta_1 + \beta_2 VWM_i^e + \beta_3 I_{1i} + \beta_4 I_{12i} + \varepsilon_i$$

where $I_{1i} = 1$ if the return was generated in January and $I_{12i} = 1$ in December. The model can be reparameterized into three cases:

$$\begin{aligned} BH_i^e &= (\beta_1 + \beta_3) + \beta_2 VWM_i^e + \varepsilon_i && \text{January} \\ BH_i^e &= (\beta_1 + \beta_4) + \beta_2 VWM_i^e + \varepsilon_i && \text{December} \\ BH_i^e &= \beta_1 + \beta_2 VWM_i^e + \varepsilon_i && \text{Otherwise} \end{aligned}$$

Dummy interactions can be used to produce models that have both different intercepts and different slopes in January and December,

$$BH_i^e = \beta_1 + \beta_2 VWM_i^e + \beta_3 I_{1i} + \beta_4 I_{12i} + \beta_5 I_{1i} VWM_i^e + \beta_6 I_{12i} VWM_i^e + \varepsilon_i.$$

If excess returns on a portfolio were nonlinearly related to returns on the market, a simple model could be specified

$$BH_i^e = \beta_1 + \beta_2 VWM_i^e + \beta_3 (VWM_i^e)^2 + \beta_4 (VWM_i^e)^3 + \varepsilon_i.$$

Dittmar (2002) proposed a similar model to explain the cross-sectional dispersion of expected returns.

3.3 Estimation

Linear regression is also known as ordinary least squares (OLS) or simply least squares. The least-squares estimator minimizes the squared distance between the fit line (or plane if there are multiple regressors) and the regressand. The parameters are estimated as the solution to

$$\min_{\beta} (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) = \min_{\beta} \sum_{i=1}^n (Y_i - \mathbf{x}_i\beta)^2. \quad (3.6)$$

First-order conditions of this optimization problem are

$$-2\mathbf{X}'(\mathbf{y} - \mathbf{X}\beta) = -2(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\beta) = -2 \sum_{i=1}^n \mathbf{x}_i(Y_i - \mathbf{x}_i\beta) = \mathbf{0} \quad (3.7)$$

and rearranging, the least-squares estimator for β can be analytically derived.

Definition 3.1 (OLS Estimator). The ordinary least-squares estimator, denoted $\hat{\beta}$, is defined

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \quad (3.8)$$

This estimator is only reasonable if $\mathbf{X}'\mathbf{X}$ is invertible, which is equivalent to the condition that $\text{rank}(\mathbf{X}) = k$. This requirement states that no column of \mathbf{X} can be exactly expressed as a combination of the $k - 1$ remaining columns and that the number of observations is at least as large as the number of regressors ($n \geq k$). This is a weak condition and is trivial to verify in most econometric software packages: using a less than full rank matrix will generate a warning or error.

Dummy variables create one further issue worthy of special attention. Suppose dummy variables corresponding to the four quarters of the year, I_{1i}, \dots, I_{4i} , are constructed from a quarterly data set of portfolio returns. Consider a simple model with a constant and all four dummies

$$R_i = \beta_1 + \beta_2 I_{1i} + \beta_3 I_{2i} + \beta_4 I_{3i} + \beta_5 I_{4i} + \varepsilon_i.$$

It is not possible to estimate this model with all four dummy variables and the constant because the constant is a perfect linear combination of the dummy variables, and so the regressor matrix would be rank deficient. The solution is to exclude either the constant or one of the dummy variables. The choice of variable to exclude makes no difference in estimation, and only the interpretation of the estimated coefficients changes. In the case where the constant is excluded, the coefficients on the dummy variables are directly interpretable as quarterly average returns. If one of the dummy variables is excluded, for example, the first quarter dummy variable, the interpretation changes. In this parameterization,

$$R_i = \beta_1 + \beta_2 I_{2i} + \beta_3 I_{3i} + \beta_4 I_{4i} + \varepsilon_i,$$

β_1 is the average return in Q1, while $\beta_1 + \beta_j$ is the average return in Qj.

It is also important that any regressor, other than the constant, be nonconstant. Suppose a regression that included the number of years since public floatation is fitted on a data set that contains only assets that have been trading for exactly 10 years. Including both this regressor and a constant results in perfect collinearity, but, more importantly, without variability in a regressor, it is impossible to determine whether changes in the regressor (years since float) results in a change in the regressand or whether the effect is simply constant across all assets. The role that that variability of regressors plays in estimating model parameters will be revisited when studying the statistical properties of $\hat{\beta}$.

The second derivative matrix of the minimization,

$$2\mathbf{X}'\mathbf{X},$$

ensures that the solution must be a minimum as long as $\mathbf{X}'\mathbf{X}$ is positive definite, which is equivalent to a condition that $\text{rank}(\mathbf{X}) = k$.

Once the regression coefficients have been estimated, it is useful to define the fit values, $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ and sample residuals $\hat{\varepsilon} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\beta}$. Rewriting the first-order condition in terms of the explanatory variables and the residuals provides insight into the numerical properties of the residuals. An equivalent first-order condition to eq. (3.7) is

$$\mathbf{X}'\hat{\varepsilon} = 0. \tag{3.9}$$

This set of linear equations is commonly referred to as the normal equations or orthogonality conditions. This set of conditions requires that $\hat{\varepsilon}$ is outside the span of the columns of \mathbf{X} . Moreover, considering the columns of \mathbf{X} separately, $\mathbf{X}'_j\hat{\varepsilon} = 0$ for all $j = 1, 2, \dots, k$. When a column contains a

constant (an intercept in the model specification), $\boldsymbol{t}'\hat{\boldsymbol{\varepsilon}} = 0$ ($\sum_{i=1}^n \hat{\varepsilon}_i = 0$), and the mean of the residuals will be exactly 0.³

The OLS estimator of the residual variance, $\hat{\sigma}^2$, can be defined.⁴

Definition 3.2 (OLS Variance Estimator). The OLS residual variance estimator, denoted $\hat{\sigma}^2$, is defined

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}' \hat{\boldsymbol{\varepsilon}}}{n - k} \quad (3.10)$$

Definition 3.3 (Standard Error of the Regression). The standard error of the regression is defined as

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} \quad (3.11)$$

The least-squares estimator has two final noteworthy properties. First, nonsingular transformations of X and non-zero scalar transformations of Y have deterministic effects on the estimated regression coefficients. Suppose \mathbf{A} is a k by k nonsingular matrix, and c is a non-zero scalar. The coefficients of a regression of cY_i on $\mathbf{x}_i \mathbf{A}$ are

$$\begin{aligned} \tilde{\beta} &= [(\mathbf{X}\mathbf{A})'(\mathbf{X}\mathbf{A})]^{-1}(\mathbf{X}\mathbf{A})'(c\mathbf{y}) \\ &= c(\mathbf{A}'\mathbf{X}'\mathbf{X}\mathbf{A})^{-1}\mathbf{A}'\mathbf{X}'\mathbf{y} \\ &= c\mathbf{A}^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'^{-1}\mathbf{A}'\mathbf{X}'\mathbf{y} \\ &= c\mathbf{A}^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= c\mathbf{A}^{-1}\hat{\beta}. \end{aligned} \quad (3.12)$$

Second, as long as the model contains a constant, the regression coefficients on all terms except the intercept are unaffected by adding an arbitrary constant to either the regressor or the regresses. Consider transforming the standard specification,

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \dots + \beta_k X_{k,i} + \varepsilon_i$$

to

$$\tilde{Y}_i = \beta_1 + \beta_2 \tilde{X}_{2,i} + \dots + \beta_k \tilde{X}_{k,i} + \varepsilon_i$$

where $\tilde{Y}_i = Y_i + c_y$ and $\tilde{X}_{j,i} = X_{j,i} + c_{x_j}$. This model is identical to

$$Y_i = \tilde{\beta}_1 + \beta_2 X_{2,i} + \dots + \beta_k X_{k,i} + \varepsilon_i$$

where $\tilde{\beta}_1 = \beta_1 + c_y - \beta_2 c_{x_2} - \dots - \beta_k c_{x_k}$.

³ \boldsymbol{t} is an n by 1 vector of 1s.

⁴The choice of $n - k$ in the denominator will be made clear once the properties of this estimator have been examined.

	Constant	VWM^e	SMB	HML	MOM	$\hat{\sigma}$
SL^e	-0.15	1.09	1.02	-0.26	-0.03	0.99
SM^e	0.08	0.96	0.82	0.35	-0.00	0.77
SH^e	0.05	1.00	0.87	0.69	-0.00	0.56
BL^e	0.12	0.99	-0.15	-0.28	-0.00	0.69
BM^e	-0.05	0.98	-0.13	0.31	-0.00	1.15
BH^e	-0.09	1.08	0.00	0.76	-0.04	1.06

Table 3.3: Estimated regression coefficients from the model $R_i^{pi} = \beta_1 + \beta_2 VWM_i^e + \beta_3 SMB_i + \beta_4 HML_i + \beta_5 MOM_i + \varepsilon_i$, where R_i^{pi} is the excess return on one of the six size and value sorted portfolios. The final column contains the standard error of the regression.

3.3.1 Estimation of Cross-Section regressions of returns on factors

Table 3.3 contains the estimated regression coefficients as well as the standard error of the regression for the six portfolios in the Fama-French data set in a specification that includes all four factors and a constant. There has been a substantial decrease in the magnitude of the standard error of the regression relative to the standard deviation of the original data. The next section will formalize how this reduction is interpreted.

3.4 Assessing Fit

Once the parameters have been estimated, the next step is to determine whether the model fits the data. The minimized sum of squared errors, the optimization's objective, is an obvious choice to assess fit. However, there is an important drawback to using the sum of squared errors: changes in the scale of Y_i alter the minimized sum of squared errors without changing the fit. It is necessary to distinguish between the portions of \mathbf{y} explained by \mathbf{X} from those that are not to construct a scale-free metric.

The projection matrix, $\mathbf{P}_\mathbf{X}$, and the annihilator matrix, $\mathbf{M}_\mathbf{X}$, are useful when decomposing the regressand into the explained component and the residual.

Definition 3.4 (Projection Matrix). The projection matrix, a symmetric idempotent matrix that produces the projection of a variable onto the space spanned by \mathbf{X} , denoted $\mathbf{P}_\mathbf{X}$, is defined

$$\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \quad (3.13)$$

Definition 3.5 (Annihilator Matrix). The annihilator matrix, a symmetric idempotent matrix that produces the projection of a variable onto the null space of \mathbf{X}' , denoted $\mathbf{M}_\mathbf{X}$, is defined

$$\mathbf{M}_\mathbf{X} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'. \quad (3.14)$$

These two matrices have some desirable properties. Both the fitted value of \mathbf{y} ($\hat{\mathbf{y}}$) and the estimated errors, $\hat{\varepsilon}$, can be expressed in terms of these matrices as $\hat{\mathbf{y}} = \mathbf{P}_\mathbf{X}\mathbf{y}$ and $\hat{\varepsilon} = \mathbf{M}_\mathbf{X}\mathbf{y}$, respectively. These

matrices are also idempotent: $\mathbf{P}_X \mathbf{P}_X = \mathbf{P}_X$ and $\mathbf{M}_X \mathbf{M}_X = \mathbf{M}_X$ and orthogonal: $\mathbf{P}_X \mathbf{M}_X = \mathbf{0}$. The projection matrix returns the portion of \mathbf{y} that lies in the linear space spanned by \mathbf{X} , while the annihilator matrix returns the portion of \mathbf{y} in the null space of \mathbf{X} . In essence, \mathbf{M}_X annihilates any portion of \mathbf{y} explainable by \mathbf{X} , leaving only the residuals.

Decomposing \mathbf{y} using the projection and annihilator matrices,

$$\mathbf{y} = \mathbf{P}_X \mathbf{y} + \mathbf{M}_X \mathbf{y}$$

which follows since $\mathbf{P}_X + \mathbf{M}_X = \mathbf{I}_n$. The squared observations can be decomposed

$$\begin{aligned}\mathbf{y}' \mathbf{y} &= (\mathbf{P}_X \mathbf{y} + \mathbf{M}_X \mathbf{y})' (\mathbf{P}_X \mathbf{y} + \mathbf{M}_X \mathbf{y}) \\ &= \mathbf{y}' \mathbf{P}_X \mathbf{P}_X \mathbf{y} + \mathbf{y}' \mathbf{P}_X \mathbf{M}_X \mathbf{y} + \mathbf{y}' \mathbf{M}_X \mathbf{P}_X \mathbf{y} + \mathbf{y}' \mathbf{M}_X \mathbf{M}_X \mathbf{y} \\ &= \mathbf{y}' \mathbf{P}_X \mathbf{y} + 0 + 0 + \mathbf{y}' \mathbf{M}_X \mathbf{y} \\ &= \mathbf{y}' \mathbf{P}_X \mathbf{y} + \mathbf{y}' \mathbf{M}_X \mathbf{y}\end{aligned}$$

noting that \mathbf{P}_X and \mathbf{M}_X are idempotent and $\mathbf{P}_X \mathbf{M}_X = \mathbf{0}_n$. These three quantities are often referred to as⁵

$$\mathbf{y}' \mathbf{y} = \sum_{i=1}^n Y_i^2 \quad \text{Uncentered Total Sum of Squares (TSS}_U\text{)} \quad (3.15)$$

$$\mathbf{y}' \mathbf{P}_X \mathbf{y} = \sum_{i=1}^n (\mathbf{x}_i \hat{\beta})^2 \quad \text{Uncentered Regression Sum of Squares (RSS}_U\text{)} \quad (3.16)$$

$$\mathbf{y}' \mathbf{M}_X \mathbf{y} = \sum_{i=1}^n (Y_i - \mathbf{x}_i \hat{\beta})^2 \quad \text{Uncentered Sum of Squared Errors (SSE}_U\text{).} \quad (3.17)$$

Dividing through by $\mathbf{y}' \mathbf{y}$

$$\frac{\mathbf{y}' \mathbf{P}_X \mathbf{y}}{\mathbf{y}' \mathbf{y}} + \frac{\mathbf{y}' \mathbf{M}_X \mathbf{y}}{\mathbf{y}' \mathbf{y}} = 1$$

or

$$\frac{\text{RSS}_U}{\text{TSS}_U} + \frac{\text{SSE}_U}{\text{TSS}_U} = 1.$$

This identity expresses the scale-free total variation in \mathbf{y} that is captured by \mathbf{X} ($\mathbf{y}' \mathbf{P}_X \mathbf{y}$) and that which is not ($\mathbf{y}' \mathbf{M}_X \mathbf{y}$). The portion of the total variation explained by \mathbf{X} is known as the uncentered R^2 (R^2_U),

⁵There is no consensus about the names of these quantities. In some texts, the component capturing the fit portion is known as the Regression Sum of Squares (RSS) while in others, it is known as the Explained Sum of Squares (ESS), while the portion attributable to the errors is known as the Sum of Squared Errors (SSE), the Sum of Squared Residuals (SSR), the Residual Sum of Squares (RSS) or the Error Sum of Squares (ESS). The choice to use SSE and RSS in this text was to ensure the reader that SSE must be the component of the squared observations relating to the error variation.

Definition 3.6 (Uncentered $R^2(R_U^2)$). The uncentered R^2 , which is used in models that do not include an intercept, is defined

$$R_U^2 = \frac{RSS_U}{TSS_U} = 1 - \frac{SSE_U}{TSS_U} \quad (3.18)$$

While R_U^2 is scale-free, it suffers from one shortcoming. Suppose a constant is added to \mathbf{y} so that the TSS_U changes to $(\mathbf{y} + c)'(\mathbf{y} + c)$. The identity still holds, and so $(\mathbf{y} + c)'(\mathbf{y} + c)$ must increase (for a sufficiently large c). In turn, one of the right-hand side variables must also grow larger. In the usual case where the model contains a constant, the increase will occur in the RSS_U ($\mathbf{y}'\mathbf{P}_{\mathbf{X}}\mathbf{y}$), and as c becomes arbitrarily large, uncentered R^2 will asymptote to one. A centered measure *computed using deviations from the mean* rather than on levels overcomes this limitation.

Let $\tilde{\mathbf{y}} = \mathbf{y} - \bar{\mathbf{y}} = \mathbf{M}_t \mathbf{y}$ where $\mathbf{M}_t = \mathbf{I}_n - t(t't)^{-1}t'$ is matrix which subtracts the mean from a vector of data. Then

$$\begin{aligned} \mathbf{y}'\mathbf{M}_t\mathbf{P}_{\mathbf{X}}\mathbf{M}_t\mathbf{y} + \mathbf{y}'\mathbf{M}_t\mathbf{M}_{\mathbf{X}}\mathbf{M}_t\mathbf{y} &= \mathbf{y}'\mathbf{M}_t\mathbf{y} \\ \frac{\mathbf{y}'\mathbf{M}_t\mathbf{P}_{\mathbf{X}}\mathbf{M}_t\mathbf{y}}{\mathbf{y}'\mathbf{M}_t\mathbf{y}} + \frac{\mathbf{y}'\mathbf{M}_t\mathbf{M}_{\mathbf{X}}\mathbf{M}_t\mathbf{y}}{\mathbf{y}'\mathbf{M}_t\mathbf{y}} &= 1 \end{aligned}$$

or more compactly

$$\frac{\tilde{\mathbf{y}}'\mathbf{P}_{\mathbf{X}}\tilde{\mathbf{y}}}{\tilde{\mathbf{y}}'\tilde{\mathbf{y}}} + \frac{\tilde{\mathbf{y}}'\mathbf{M}_{\mathbf{X}}\tilde{\mathbf{y}}}{\tilde{\mathbf{y}}'\tilde{\mathbf{y}}} = 1.$$

Centered R^2 (R_C^2) is defined analogously to uncentered replacing the uncentered sums of squares with their centered counterparts.

Definition 3.7 (Centered $R^2(R_C^2)$). The uncentered R^2 , used in models that include an intercept, is defined

$$R_C^2 = \frac{RSS_C}{TSS_C} = 1 - \frac{SSE_C}{TSS_C} \quad (3.19)$$

where

$$\mathbf{y}'\mathbf{M}_t\mathbf{y} = \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad \text{Centered Total Sum of Squares (TSS}_C\text{)} \quad (3.20)$$

$$\mathbf{y}'\mathbf{M}_t\mathbf{P}_{\mathbf{X}}\mathbf{M}_t\mathbf{y} = \sum_{i=1}^n (\mathbf{x}_i\hat{\beta} - \bar{\mathbf{x}}\hat{\beta})^2 \quad \text{Centered Regression Sum of Squares (RSS}_C\text{)} \quad (3.21)$$

$$\mathbf{y}'\mathbf{M}_t\mathbf{M}_{\mathbf{X}}\mathbf{M}_t\mathbf{y} = \sum_{i=1}^n (Y_i - \mathbf{x}_i\hat{\beta})^2 \quad \text{Centered Sum of Squared Errors (SSE}_C\text{).} \quad (3.22)$$

and where $\bar{\mathbf{x}} = n^{-1} \sum_{i=1}^n \mathbf{x}_i$.

The expressions R^2 , SSE, RSS, and TSS should be assumed to correspond to the centered version unless further qualified. With two versions of R^2 available that generally differ, which should be used? Centered should be used if the model is centered (contains a constant), and uncentered should

be used when it does not. Failing to select the correct R^2 can lead to incorrect conclusions about the model's fit, and mixing the definitions can lead to a nonsensical R^2 that falls outside of $[0, 1]$. For instance, computing R^2 using the centered version when the model does not contain a constant often results in a negative value when

$$R^2 = 1 - \frac{SSE_C}{TSS_C}.$$

Most software will return centered R^2 , and caution is warranted if a model is fit without a constant.

R^2 does have some caveats. First, adding an additional regressor will always (weakly) increase the R^2 since the sum of squared errors cannot increase by the inclusion of an additional regressor. This renders R^2 useless in discriminating between two models where one is nested within the other. One solution to this problem is to use the degree of freedom adjusted R^2 .

Definition 3.8 (Adjusted R^2 (\bar{R}^2)). The adjusted R^2 , which adjusts for the number of estimated parameters, is defined

$$\bar{R}^2 = 1 - \frac{\frac{SSE}{n-k}}{\frac{TSS}{n-1}} = 1 - \frac{SSE}{TSS} \frac{n-1}{n-k}. \quad (3.23)$$

\bar{R}^2 will increase if the reduction in the SSE is large enough to compensate for a loss of one degree of freedom, captured by the $n - k$ term. However, if the SSE does not change, \bar{R}^2 will decrease. \bar{R}^2 is preferable to R^2 for comparing models, although the topic of model selection will be more formally considered at the end of this chapter. \bar{R}^2 , like R^2 , should be constructed from the appropriate versions of the RSS, SSE, and TSS (either centered or uncentered).

Second, R^2 is not invariant to changes in the regressand. A frequent mistake is to use R^2 to compare the fit from two models with different regressands, for instance, Y_i and $\ln(Y_i)$. These numbers are incomparable, and this type of comparison must be avoided. Moreover, R^2 is even sensitive to more benign transformations. Suppose a simple model is postulated,

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i,$$

and a model logically consistent with the original model,

$$Y_i - X_i = \beta_1 + (\beta_2 - 1)X_i + \varepsilon_i,$$

is estimated. The R^2 's from these models will generally differ. For example, suppose the original coefficient on x_i was 1. Subtracting x_i will reduce the explanatory power of x_i to 0, rendering it useless and resulting in a R^2 of 0 irrespective of the R^2 in the original model.

3.4.1 Example: R^2 and \bar{R}^2 in Cross-Sectional Factor models

To illustrate the use of R^2 , consider alternative models of BH^e that include one or more risk factors. The R^2 values in the top half of Table 3.4 show that R^2 never declines as additional variables are added. Note that the adjusted measure of fit, \bar{R}_U^2 , also never declines, although it grows more slowly. The monotonic pattern occurs since the adjustment penalty is small when the sample size n is large, as is the case here. The table only shows the correct version of the R^2 – centered for models that contain a constant and uncentered for those that do not.

Regressand	Regressors	R_U^2	\bar{R}_U^2	R_C^2	\bar{R}_C^2
BH^e	1, VME^e	0.7620	0.7616	—	—
BH^e	1, VME^e, SMB	0.7644	0.7637	—	—
BH^e	1, VME^e, SMB, HML	0.9535	0.9533	—	—
BH^e	1, VME^e, SMB, HML, MOM	0.9543	0.9541	—	—
BH^e	VWM^e	—	—	0.7656	0.7653
$10 + BH^e$	1, VME^e	0.7620	0.7616	—	—
$10 + BH^e$	VME^e	—	—	0.2275	0.2264
$10 \times BH^e$	1, VME^e	0.7620	0.7616	—	—
$10 \times BH^e$	VME^e	—	—	0.7656	0.7653
$BH^e - VME^e$	1, VME^e	0.0024	0.0009	—	—
$\sum_Y BH^e$	1, $\sum_Y VME^e$	0.6800	0.6743	—	—

Table 3.4: Centered and uncentered R^2 and \bar{R}^2 from models with regressor or regressand changes. Only the correct version of the R^2 is shown – centered for models that contain a constant as indicated by 1 in the regressor list, or uncentered for models that do not. The top rows demonstrate how R^2 and its adjusted version change as additional variables are added. The bottom two rows demonstrate how changes in the regressand – the left-hand-side variable – affect the R^2 .

The bottom half of the table shows how R^2 changes when the regressand changes. The R^2 in models that include a constant are invariant to constant shifts in the regressand. The R_U^2 of the model that regresses $10 + BH^e$ on a constant and the excess market is identical to the same model only using BH^e . This relationship does not hold for models that do not contain a constant and R_C^2 changes when 10 is added to the return. Both measures are invariant to multiplicative adjustments. The penultimate line shows that R^2 is *not* invariant to changes in the regressand that do not fundamentally alter the interpretation of the model. In this model, the difference in returns, $BH^e - VMW^e$, is regressed on a constant and the excess market, $\hat{\gamma}_2$, in this model

$$BH_i^e - VWM^e = \gamma_1 + \gamma_2 VWM_i^e + \varepsilon_i.$$

will be *exactly* 1 less than the coefficient in the model

$$BH_i^e = \beta_1 + \beta_2 VWM_i^e + \varepsilon_i.$$

While these two models are conceptually identical and describe the same relationship between BH^e , the R^2 changes. In this example, the coefficient on VWM^e is near zero since the coefficient in the original specification is near 1. The R^2 of the return difference is near 0 even though the market is an important determinant of the the Big-High portfolio's return. The final line shows the regression coefficient of the annual return of BH^e ($\sum_Y BH^e$) on the annual return on the market ($\sum_Y VWM^e$). This type of aggregation also changes the R^2 . These final two results highlight a common form of misuse of R^2 : do not compare the values of R^2 in models with different regressands.

3.5 Assumptions

Thus far, all of the derivations and identities presented are purely numerical. They do not indicate whether $\hat{\beta}$ is a reasonable way to estimate β . It is necessary to make some assumptions about the innovations and the regressors to provide a statistical interpretation of $\hat{\beta}$. Two broad classes of assumptions can be used to analyze the behavior of $\hat{\beta}$: the classical framework (also known as the small-sample or finite-sample framework) and asymptotic analysis (also known as the large-sample framework).

Neither method is ideal. The small-sample framework is precise in that the exact distribution of regressors and test statistics are known. This precision comes at the cost of many restrictive assumptions – assumptions not usually plausible in financial applications. On the other hand, asymptotic analysis requires few restrictive assumptions and is broadly applicable to financial data, although the results are only exact if the number of observations is infinite. Asymptotic analysis is still useful for examining the behavior in finite samples when the sample size is large enough for the asymptotic distribution to approximate the finite-sample distribution reasonably well.

This leads to the most important question of asymptotic analysis: How large does n need to be before the approximation is reasonable? Unfortunately, the answer to this question is “*it depends*”. In simple cases, where residuals are independent and identically distributed, as few as 30 observations may be sufficient for the asymptotic distribution to be a good approximation to the finite-sample distribution. In more complex cases, anywhere from 100 to 1,000 may be needed, while in the extreme cases, where the data is heterogeneous and highly dependent, an asymptotic approximation may be poor with more than 1,000,000 observations.

The properties of $\hat{\beta}$ will be examined under both sets of assumptions. While the small-sample results are not generally applicable, it is important to understand these results as the *lingua franca* of econometrics, as well as the limitations of tests based on the classical assumptions, and to be able to detect when a test statistic may not have the intended asymptotic distribution. Six assumptions are required to examine the finite-sample distribution of $\hat{\beta}$ and establish the optimality of the OLS procedure (although many properties only require a subset).

Assumption 3.1 (Linearity). $Y_i = \mathbf{x}_i\beta + \varepsilon_i$

This assumption states the obvious condition necessary for least squares to be a reasonable method to estimate the β . It further imposes a less obvious condition, that \mathbf{x}_i must be observed and measured without error. Many applications in financial econometrics include *latent* variables. Linear regression is not applicable in these cases and a more sophisticated estimator is required. In other applications, the *true* value of $x_{k,i}$ is not observed and a noisy proxy must be used, so that $\tilde{x}_{k,i} = x_{k,i} + v_{k,i}$ where $v_{k,i}$ is an error uncorrelated with $x_{k,i}$. When this occurs, ordinary least-squares estimators are misleading and a modified procedure (two-stage least squares (2SLS) or instrumental variable regression (IV)) must be used.

Assumption 3.2 (Conditional Mean). $E[\varepsilon_i | \mathbf{X}] = 0, \quad i = 1, 2, \dots, n$

This assumption states that the mean of each ε_i is zero given any $X_{k,i}$, any function of any $X_{k,i}$ or combinations of these. It is stronger than the assumption used in the asymptotic analysis and is not valid in many applications (e.g., time-series data). When the regressand and regressor consist of time-series data, this assumption may be violated and $E[\varepsilon_i | \mathbf{x}_{i+j}] \neq 0$ for some j . This assumption also implies that the correct form of $X_{k,i}$ enters the regression, that $E[\varepsilon_i] = 0$ (through a simple application

of the law of iterated expectations), and that the innovations are uncorrelated with the regressors, so that $E[\varepsilon_{i'}x_{j,i}] = 0, i' = 1, 2, \dots, n, i = 1, 2, \dots, n, j = 1, 2, \dots, k$.

Assumption 3.3 (Rank). *The rank of \mathbf{X} is k with probability 1.*

This assumption is needed to ensure that $\hat{\beta}$ is identified and can be estimated. In practice, it requires that no regressor is perfectly co-linear with the others, that the number of observations is at least as large as the number of regressors ($n \geq k$) and that variables other than a constant have non-zero variance.

Assumption 3.4 (Conditional Homoskedasticity). $V[\varepsilon_i|\mathbf{X}] = \sigma^2$

Homoskedasticity is rooted in *homo* (same) and *skedannumi* (scattering) and in modern English means that the residuals have identical variances. This assumption is required to establish the optimality of the OLS estimator and it specifically rules out the case where the variance of an innovation is a function of a regressor.

Assumption 3.5 (Conditional Correlation). $E[\varepsilon_i\varepsilon_j|\mathbf{X}] = 0, i = 1, 2, \dots, n, j = i+1, \dots, n$

Assuming the residuals are conditionally uncorrelated is convenient when coupled with the homoskedasticity assumption, and the residuals covariance is $\sigma^2\mathbf{I}_n$. Like homoskedasticity, this assumption is needed for establishing the optimality of the least-squares estimator.

Assumption 3.6 (Conditional Normality). $\varepsilon|\mathbf{X} \sim N(\mathbf{0}, \Sigma)$

Assuming a specific distribution is very restrictive – results based on this assumption will only be correct if the errors are actually normal – but this assumption allows for precise statements about the finite-sample distribution of $\hat{\beta}$ and test statistics. This assumption, when combined with assumptions 3.4 and 3.5, provides a simple distribution for the innovations: $\varepsilon_i|\mathbf{X} \xrightarrow{d} N(0, \sigma^2)$.

3.6 Small-Sample Properties of OLS estimators

Using these assumptions, many useful properties of $\hat{\beta}$ can be derived. Recall that $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

Theorem 3.1 (Bias of $\hat{\beta}$). *Under assumptions 3.1 - 3.3*

$$E[\hat{\beta}|\mathbf{X}] = \beta. \quad (3.24)$$

While unbiasedness is a desirable property, it is not particularly meaningful without further qualification. For instance, an estimator which is unbiased, but does not increase in precision as the sample size increases is generally not desirable. Fortunately, $\hat{\beta}$ is not only unbiased, it has a variance that goes to zero.

Theorem 3.2 (Variance of $\hat{\beta}$). *Under assumptions 3.1 - 3.5*

$$V[\hat{\beta}|\mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \quad (3.25)$$

Under the conditions necessary for unbiasedness for $\hat{\beta}$, plus assumptions about homoskedasticity and the conditional correlation of the residuals, the form of the variance is simple. Consistency follows since

$$\begin{aligned} (\mathbf{X}'\mathbf{X})^{-1} &= \left(n \frac{\sum_{i=1}^n \mathbf{x}_i' \mathbf{x}_i}{n} \right)^{-1} \\ &\approx \frac{1}{n} \mathbf{E} [\mathbf{x}_i' \mathbf{x}_i]^{-1} \end{aligned} \quad (3.26)$$

will be declining as the sample size increases.

However, $\hat{\beta}$ has an even stronger property under the same assumptions. It is BLUE: *Best Linear Unbiased Estimator*. Best, in this context, means that it has the lowest variance among all other linear unbiased estimators. While this is a strong result, a few words of caution are needed to properly interpret this result. The class of Linear Unbiased Estimators (LUEs) is small in the universe of all unbiased estimators. Saying OLS is the “best” is akin to a one-armed boxer claiming to be the best one-arm boxer. While possibly true, she probably would not stand a chance against a two-armed opponent.

Theorem 3.3 (Gauss-Markov Theorem). *Under assumptions 3.1-3.5, $\hat{\beta}$ is the minimum variance estimator among all linear unbiased estimators. That is $\mathbf{V}[\tilde{\beta}|\mathbf{X}] - \mathbf{V}[\hat{\beta}|\mathbf{X}]$ is positive semi-definite where $\tilde{\beta} = \mathbf{C}\mathbf{y}$, $\mathbf{E}[\tilde{\beta}] = \beta$ and $\mathbf{C} \neq (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.*

Letting $\tilde{\beta}$ be any other *linear, unbiased* estimator of β , it must have a larger covariance. However, many estimators, including most maximum likelihood estimators, are nonlinear and so are not necessarily less efficient. Finally, making use of the normality assumption, it is possible to determine the conditional distribution of $\hat{\beta}$.

Theorem 3.4 (Distribution of $\hat{\beta}$). *Under assumptions 3.1 – 3.6,*

$$\hat{\beta}|\mathbf{X} \sim N(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}) \quad (3.27)$$

Theorem 3.4 should not be surprising. $\hat{\beta}$ is a linear combination of (jointly) normally distributed random variables and thus is also normally distributed. Normality is also useful for establishing the relationship between the estimated residuals $\hat{\epsilon}$ and the estimated parameters $\hat{\beta}$.

Theorem 3.5 (Conditional Independence of $\hat{\epsilon}$ and $\hat{\beta}$). *Under assumptions 3.1 - 3.6, $\hat{\epsilon}$ is independent of $\hat{\beta}$, conditional on \mathbf{X} .*

One implication of this theorem is that $\text{Cov}(\hat{\epsilon}_i, \hat{\beta}_j | \mathbf{X}) = 0$ $i = 1, 2, \dots, n$, $j = 1, 2, \dots, k$. As a result, functions of $\hat{\epsilon}$ will be independent of functions of $\hat{\beta}$, a property useful in deriving distributions of test statistics that depend on both. Finally, in the small-sample setup, the exact distribution of the sample error variance estimator, $\hat{\sigma}^2 = \hat{\epsilon}'\hat{\epsilon}/(n-k)$, can be derived.

Theorem 3.6 (Distribution of $\hat{\sigma}^2$).

$$(n-k) \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-k}^2$$

where $\hat{\sigma}^2 = \frac{\mathbf{y}'\mathbf{M}_{\mathbf{X}}\mathbf{y}}{n-k} = \frac{\hat{\epsilon}'\hat{\epsilon}}{n-k}$.

Since $\hat{\epsilon}_i$ is a normal random variable, once it is standardized and squared, it should be a χ_1^2 . The change in the divisor from n to $n-k$ reflects the loss in degrees of freedom due to the k estimated parameters.

3.7 Maximum Likelihood

Once the assumption that the innovations are conditionally normal has been made, conditional maximum likelihood is an obvious method to estimate the unknown parameters (β, σ^2) . Conditioning on \mathbf{X} , and assuming the innovations are normal, homoskedastic, and conditionally uncorrelated, the likelihood is given by

$$f(\mathbf{y}|\mathbf{X}; \beta, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)}{2\sigma^2}\right) \quad (3.28)$$

and, taking logs, the log likelihood

$$l(\beta, \sigma^2; \mathbf{y}|\mathbf{X}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)}{2\sigma^2}. \quad (3.29)$$

Recall that the logarithm is a monotonic, strictly increasing transformation, and the extremum points of the log-likelihood and the likelihood will occur at the same parameters. Maximizing the likelihood with respect to the unknown parameters, there are $k + 1$ first-order conditions

$$\frac{\partial l(\beta, \sigma^2; \mathbf{y}|\mathbf{X})}{\partial \beta} = \frac{\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\beta})}{\sigma^2} = 0 \quad (3.30)$$

$$\frac{\partial l(\beta, \sigma^2; \mathbf{y}|\mathbf{X})}{\partial \hat{\sigma}^2} = -\frac{n}{2\hat{\sigma}^2} + \frac{(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta})}{2\hat{\sigma}^4} = 0. \quad (3.31)$$

The first set of conditions is identical to the first-order conditions of the least-squares estimator ignoring the scaling by σ^2 , assumed to be greater than 0. The solution is

$$\hat{\beta}^{\text{MLE}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (3.32)$$

$$\hat{\sigma}^{2\text{MLE}} = n^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta}) = n^{-1}\hat{\varepsilon}'\hat{\varepsilon}. \quad (3.33)$$

The regression coefficients are identical under maximum likelihood and OLS, although the divisor in $\hat{\sigma}^2$ and $\hat{\sigma}^{2\text{MLE}}$ differ.

It is important to note that the derivation of the OLS estimator does not require an assumption of normality. Moreover, the unbiasedness, variance, and BLUE properties do not rely on the conditional normality of residuals. However, if the innovations are homoskedastic, uncorrelated and normal, the results of the Gauss-Markov theorem can be strengthened using the Cramer-Rao lower bound.

Theorem 3.7 (Cramer-Rao Inequality). *Let $f(\mathbf{z}; \theta)$ be the joint density of \mathbf{z} where θ is a k dimensional parameter vector. Let $\hat{\theta}$ be an unbiased estimator of θ_0 with finite covariance. Under some regularity condition on $f(\cdot)$*

$$\mathbf{V}[\hat{\theta}] \geq \mathcal{I}^{-1}(\theta_0)$$

where

$$\mathcal{I} = -E\left[\frac{\partial^2 \ln f(\mathbf{z}; \theta)}{\partial \theta \partial \theta'}\Big|_{\theta=\theta_0}\right] \quad (3.34)$$

and

$$\mathcal{J} = E \left[\frac{\partial \ln f(\mathbf{z}; \theta)}{\partial \theta} \frac{\partial \ln f(\mathbf{z}; \theta)}{\partial \theta'} \Big|_{\theta=\theta_0} \right] \quad (3.35)$$

and, under some additional regularity conditions,

$$\mathcal{I}(\theta_0) = \mathcal{J}(\theta_0).$$

The last part of this theorem is the information matrix equality (IME) and when a model is correctly specified in its entirety, the expected covariance of the scores is equal to negative of the expected hessian.⁶ The IME will be revisited in later chapters. The second order conditions,

$$\frac{\partial^2 l(\beta, \sigma^2; \mathbf{y} | \mathbf{X})}{\partial \beta \partial \beta'} = -\frac{\mathbf{X}' \mathbf{X}}{\hat{\sigma}^2} \quad (3.36)$$

$$\frac{\partial^2 l(\beta, \sigma^2; \mathbf{y} | \mathbf{X})}{\partial \beta \partial \sigma^2} = -\frac{\mathbf{X}' (\mathbf{y} - \mathbf{X}\beta)}{\sigma^4} \quad (3.37)$$

$$\frac{\partial^2 l(\beta, \sigma^2; \mathbf{y} | \mathbf{X})}{\partial^2 \sigma^2} = \frac{n}{2\sigma^4} - \frac{(\mathbf{y} - \mathbf{X}\beta)' (\mathbf{y} - \mathbf{X}\beta)}{\sigma^6} \quad (3.38)$$

are needed to find the lower bound for the covariance of the estimators of β and σ^2 . Taking expectations of the second derivatives,

$$E \left[\frac{\partial^2 l(\beta, \sigma^2; \mathbf{y} | \mathbf{X})}{\partial \beta \partial \beta'} \right] = -\frac{\mathbf{X}' \mathbf{X}}{\sigma^2} \quad (3.39)$$

$$E \left[\frac{\partial^2 l(\beta, \sigma^2; \mathbf{y} | \mathbf{X})}{\partial \beta \partial \sigma^2} \right] = \mathbf{0} \quad (3.40)$$

$$E \left[\frac{\partial^2 l(\beta, \sigma^2; \mathbf{y} | \mathbf{X})}{\partial^2 \sigma^2} \right] = -\frac{n}{2\sigma^4} \quad (3.41)$$

and so the lower bound for the variance of $\hat{\beta} = \hat{\beta}^{MLE}$ is $\sigma^2(\mathbf{X}' \mathbf{X})^{-1}$. Theorem 3.2 show that $\sigma^2(\mathbf{X}' \mathbf{X})^{-1}$ is also the variance of the OLS estimator $\hat{\beta}$ and so the Gauss-Markov theorem can be strengthened in the case of conditionally homoskedastic, uncorrelated normal residuals.

Theorem 3.8 (Best Unbiased Estimator). *Under assumptions 3.1 - 3.6, $\hat{\beta} = \hat{\beta}^{MLE}$ is the best unbiased estimator of β .*

The difference between this theorem and the Gauss-Markov theorem is subtle but important. The class of estimators is no longer restricted to include only linear estimators and so this result is both broad and powerful: MLE (or OLS) is an ideal estimator under these assumptions (in the sense that no other unbiased estimator, linear or not, has a lower variance). This results does not extend to the variance estimator since $E[\hat{\sigma}^{2MLE}] = \frac{n}{n-k} \sigma^2 \neq \sigma^2$, and so the optimality of $\hat{\sigma}^{2MLE}$ cannot be established using the Cramer-Rao theorem.

⁶There are quite a few regularity conditions for the IME to hold, but discussion of these is beyond the scope of this course. Interested readers should see White (1996) for a thorough discussion.

3.8 Small-Sample Hypothesis Testing

Most regressions are estimated to test implications of economic or finance theory. Hypothesis testing is the mechanism used to determine whether data and theory are congruent. Formalized in terms of β , the null hypothesis (also known as the maintained hypothesis) is formulated as

$$H_0 : \mathbf{R}(\beta) - \mathbf{r} = \mathbf{0} \quad (3.42)$$

where $\mathbf{R}(\cdot)$ is a function from \mathbb{R}^k to \mathbb{R}^m , $m \leq k$ and \mathbf{r} is an m by 1 vector. Initially, a subset of all hypotheses, those in the linear equality hypotheses class, formulated

$$H_0 : \mathbf{R}\beta - \mathbf{r} = \mathbf{0} \quad (3.43)$$

will be examined where \mathbf{R} is a m by k matrix. In subsequent chapters, more general test specifications including nonlinear restrictions on the parameters will be considered. All hypotheses in this class can be written as weighted sums of the regression coefficients,

$$\begin{aligned} R_{11}\beta_1 + R_{12}\beta_2 \dots + R_{1k}\beta_k &= r_1 \\ R_{21}\beta_1 + R_{22}\beta_2 \dots + R_{2k}\beta_k &= r_2 \\ &\vdots \\ R_{m1}\beta_1 + R_{m2}\beta_2 \dots + R_{mk}\beta_k &= r_i \end{aligned}$$

Each constraint is represented as a row in the above set of equations. Linear equality constraints can be used to test parameter restrictions such as

$$\begin{aligned} \beta_1 &= 0 \\ 3\beta_2 + \beta_3 &= 1 \\ \sum_{j=1}^k \beta_j &= 0 \\ \beta_1 = \beta_2 = \beta_3 &= 0. \end{aligned} \quad (3.44)$$

For instance, if the unrestricted model is

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \beta_5 X_{5,i} + \varepsilon_i$$

the hypotheses in eq. (3.44) can be described in terms of \mathbf{R} and \mathbf{r} as

H_0	\mathbf{R}	\mathbf{r}
$\beta_1 = 0$	$[1 \ 0 \ 0 \ 0 \ 0]$	0
$3\beta_2 + \beta_3 = 1$	$[0 \ 3 \ 1 \ 0 \ 0]$	1
$\sum_{j=1}^k \beta_j = 0$	$[0 \ 1 \ 1 \ 1 \ 1]$	0
$\beta_1 = \beta_2 = \beta_3 = 0$	$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$

When using linear equality constraints, alternatives are specified as $H_1 : \mathbf{R}\beta - \mathbf{r} \neq 0$. Once both the null and the alternative hypotheses have been postulated, it is necessary to discern whether the data are consistent with the null hypothesis. Three classes of statistics will be described to test these hypotheses: Wald, Lagrange Multiplier and Likelihood Ratio. Wald tests are perhaps the most intuitive: they directly test whether $\mathbf{R}\beta - \mathbf{r}$ is close to zero. Lagrange Multiplier tests incorporate the constraint into the least-squares problem using a Lagrangian. If the constraint has a small effect on the minimized sum of squares, the Lagrange multipliers, often described as the shadow price of the constraint in economic applications, should be close to zero. The magnitude of these forms the basis of the LM test statistic. Finally, likelihood ratios test whether the data are less likely under the null than they are under the alternative. If the null hypothesis is not restrictive this ratio should be close to one and the difference in the log-likelihoods should be small.

3.8.1 t -tests

T-tests can be used to test a single hypothesis involving one or more coefficients,

$$H_0 : \mathbf{R}\beta = r$$

where \mathbf{R} is a 1 by k vector and r is a scalar. Recall from theorem 3.4, $\hat{\beta} - \beta \sim N(\mathbf{0}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$. Under the null, $\mathbf{R}(\hat{\beta} - \beta) = \mathbf{R}\hat{\beta} - \mathbf{R}\beta = \mathbf{R}\hat{\beta} - r$ and applying the properties of normal random variables,

$$\mathbf{R}\hat{\beta} - r \sim N(\mathbf{0}, \sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}').$$

A simple test can be constructed

$$z = \frac{\mathbf{R}\hat{\beta} - r}{\sqrt{\sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}}}, \quad (3.45)$$

where $z \sim N(0, 1)$. To perform a test with size α , the value of z can be compared to the critical values of the standard normal and rejected if $|z| > C_\alpha$ where C_α is the $1 - \alpha$ quantile of a standard normal. However, z is an infeasible statistic since it depends on an unknown quantity, σ^2 . The natural solution is to replace the unknown parameter with an estimate. Dividing z by $\sqrt{\frac{s^2}{\sigma^2}}$ and simplifying,

$$\begin{aligned} t &= \frac{z}{\sqrt{\frac{s^2}{\sigma^2}}} \\ &= \frac{\mathbf{R}\hat{\beta} - r}{\sqrt{\sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}'}} \sqrt{\frac{s^2}{\sigma^2}} \\ &= \frac{\mathbf{R}\hat{\beta} - r}{\sqrt{s^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}'}}. \end{aligned} \quad (3.46)$$

Note that the denominator $(n - k) \frac{s^2}{\sigma^2} \sim \chi_{n-k}^2$, and so t is the ratio of a standard normal to the square root of a χ_v^2 normalized by its standard deviation. As long as the standard normal in the numerator and the χ_v^2 are independent, this ratio will have a Student's t distribution.

Definition 3.9 (Student's t distribution). Let $z \sim N(0, 1)$ (standard normal) and let $w \sim \chi_v^2$ where z and w are independent. Then

$$\frac{z}{\sqrt{\frac{w}{v}}} \sim t_v. \quad (3.47)$$

2

The independence of $\hat{\beta}$ and s^2 – which is only a function of $\hat{\epsilon}$ – follows from 3.5, and so t has a Student's t distribution.

Theorem 3.9 (t -test). *Under assumptions 3.1 - 3.6,*

$$\frac{\mathbf{R}\hat{\beta} - r}{\sqrt{s^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}'}} \sim t_{n-k}. \quad (3.48)$$

As $v \rightarrow \infty$, the Student's t distribution converges to a standard normal. As a practical matter, when $v > 30$, the T distribution is close to a normal. While any single linear restriction can be tested with a t -test, the expression t -stat has become synonymous with a specific null hypothesis.

Definition 3.10 (t -stat). The t -stat of a coefficient, β_k , is the t -test value of a test of the null $H_0 : \beta_k = 0$ against the alternative $H_1 : \beta_k \neq 0$, and is computed

$$\frac{\hat{\beta}_k}{\sqrt{s^2 (\mathbf{X}'\mathbf{X})_{[kk]}^{-1}}} \quad (3.49)$$

where $(\mathbf{X}'\mathbf{X})_{[kk]}^{-1}$ is the k^{th} diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$.

The previous examples were all two-sided; the null would be rejected if the parameters differed in either direction from the null hypothesis. The T-test is also unique among these three main classes of test statistics in that it can easily be applied against both one-sided alternatives and two-sided alternatives.⁷

However, there is often a good argument to test a one-sided alternative. For instance, in tests of the market premium, theory indicates that it must be positive to induce investment. Thus, when testing the null hypothesis that a risk premium is zero, a two-sided alternative could reject in cases which are not theoretically interesting. More importantly, a one-sided alternative, when appropriate, will have more power than a two-sided alternative since the direction information in the null hypothesis can be used to tighten confidence intervals. The two types of tests involving a one-sided hypothesis are upper tail tests which test nulls of the form $H_0 : \mathbf{R}\beta \leq r$ against alternatives of the form $H_1 : \mathbf{R}\beta > r$, and lower tail tests which test $H_0 : \mathbf{R}\beta \geq r$ against $H_1 : \mathbf{R}\beta < r$.

Figure 3.1 contains the rejection regions of a t_{10} distribution. The dark gray region corresponds to the rejection region of a two-sided alternative to the null that $H_0 : \hat{\beta} = \beta^0$ for a 10% test. The light gray region, combined with the upper dark gray region corresponds to the rejection region of a one-sided upper tail test, and so test statistic between 1.372 and 1.812 would be rejected using a one-sided alternative but not with a two-sided one.

Algorithm 3.1 (t -test).

⁷Wald, LM, and LR tests can be implemented against one-sided alternatives with considerably more effort.

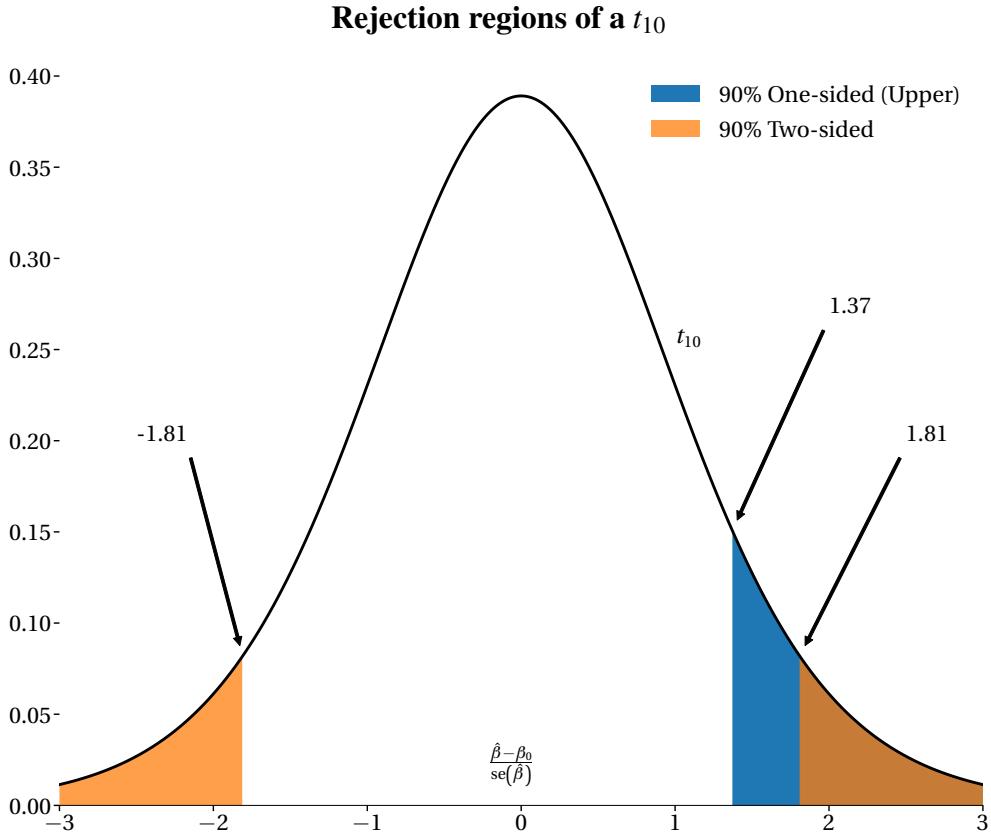


Figure 3.1: Rejection region for a t -test of the nulls $H_0 : \beta = \beta^0$ (two-sided) and $H_0 : \beta \leq \beta^0$. The two-sided rejection region is indicated by dark gray while the one-sided (upper) rejection region includes both the light and dark gray areas in the right tail.

1. Estimate $\hat{\beta}$ using least squares.
2. Compute $s^2 = (n - k)^{-1} \sum_{i=1}^n \hat{\varepsilon}_i^2$ and $s^2(\mathbf{X}'\mathbf{X})^{-1}$.
3. Construct the restriction matrix, \mathbf{R} , and the value of the restriction, r from the null hypothesis.
4. Compute $t = \frac{\mathbf{R}\hat{\beta} - r}{\sqrt{s^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'}}$.
5. Compare t to the critical value, C_α , of the t_{n-k} distribution for a test size with α . In the case of a two tailed test, reject the null hypothesis if $|t| > F_{t_v}(1 - \alpha/2)$ where $F_{t_v}(\cdot)$ is the CDF of a t_v -distributed random variable. In the case of a one-sided upper-tail test, reject if $t > F_{t_v}(1 - \alpha)$ or in the case of a one-sided lower-tail test, reject if $t < F_{t_v}(\alpha)$.

3.8.2 Wald Tests

Wald test directly examines the distance between $\mathbf{R}\beta$ and \mathbf{r} . Intuitively, if the null hypothesis is true, then $\mathbf{R}\beta - \mathbf{r} \approx 0$. In the small-sample framework, the distribution of $\mathbf{R}\beta - \mathbf{r}$ follows directly from the properties of normal random variables. Specifically,

$$\mathbf{R}\beta - \mathbf{r} \sim N(\mathbf{0}, \sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')$$

Thus, to test the null $H_0 : \mathbf{R}\beta - \mathbf{r} = 0$ against the alternative $H_0 : \mathbf{R}\beta - \mathbf{r} \neq 0$, a test statistic can be based on

$$W_{\text{Infeasible}} = \frac{(\mathbf{R}\beta - \mathbf{r})' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} (\mathbf{R}\beta - \mathbf{r})}{\sigma^2} \quad (3.50)$$

which has a χ_m^2 distribution.⁸ However, this statistic depends on an unknown quantity, σ^2 , and to operationalize W , σ^2 must be replaced with an estimate, s^2 .

$$W = \frac{(\mathbf{R}\beta - \mathbf{r})' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} (\mathbf{R}\beta - \mathbf{r})/m}{s^2} \frac{\sigma^2}{s^2} = \frac{(\mathbf{R}\beta - \mathbf{r})' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} (\mathbf{R}\beta - \mathbf{r})/m}{s^2} \quad (3.51)$$

The replacement of σ^2 with s^2 has an effect on the distribution of the estimator which follows from the definition of an F distribution.

Definition 3.11 (F distribution). Let $z_1 \sim \chi_{v_1}^2$ and let $z_2 \sim \chi_{v_2}^2$ where z_1 and z_2 are independent. Then

$$\frac{\frac{z_1}{v_1}}{\frac{z_2}{v_2}} \sim F_{v_1, v_2} \quad (3.52)$$

The conclusion that W has a $F_{m,n-k}$ distribution follows from the independence of $\hat{\beta}$ and $\hat{\epsilon}$, which in turn implies the independence of $\hat{\beta}$ and s^2 .

Theorem 3.10 (Wald test). *Under assumptions 3.1 - 3.6,*

$$\frac{(\mathbf{R}\beta - \mathbf{r})' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} (\mathbf{R}\beta - \mathbf{r})/m}{s^2} \sim F_{m,n-k} \quad (3.53)$$

Analogous to the t_v distribution, an F_{v_1, v_2} distribution converges to a scaled χ^2 in large samples ($\chi_{v_1}^2/v_1$ as $v_2 \rightarrow \infty$). Figure 3.2 contains *failure to reject* (FTR) regions for some hypothetical Wald tests. The shape of the region depends crucially on the correlation between the hypotheses being tested. For instance, panel (a) corresponds to testing a joint hypothesis where the tests are independent and have the same variance. In this case, the FTR region is a circle. Panel (d) shows the FTR region for highly correlated tests where one restriction has a larger variance.

Once W has been computed, the test statistic should be compared to the critical value of an $F_{m,n-k}$ and rejected if the test statistic is larger. Figure 3.3 contains the pdf of an $F_{5,30}$ distribution. Any $W > 2.049$ would lead to rejection of the null hypothesis using a 10% test.

The Wald test has a more common expression in terms of the SSE from both the restricted and unrestricted models. Specifically,

⁸The distribution can be derived noting that $[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-\frac{1}{2}} (\mathbf{R}\beta - \mathbf{r}) \sim N\left(0, \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\right)$ where the matrix square root makes use of a generalized inverse. A more complete discussion of reduced rank normals and generalized inverses is beyond the scope of this course.

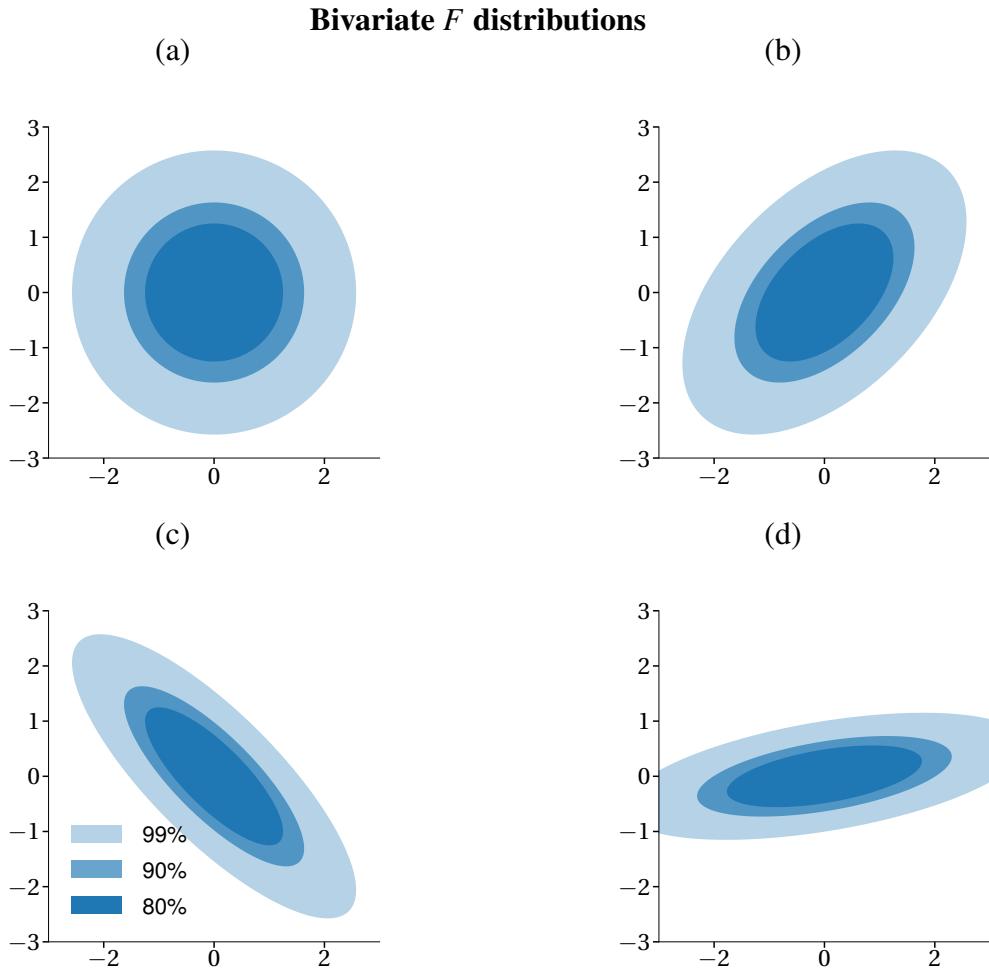


Figure 3.2: Bivariate plot of an F distribution. The four panels contain the failure-to-reject regions corresponding to 20, 10 and 1% tests. Panel (a) contains the region for uncorrelated tests. Panel (b) contains the region for tests with the same variance but a correlation of 0.5. Panel (c) contains the region for tests with a correlation of -.8 and panel (d) contains the region for tests with a correlation of 0.5 but with variances of 2 and 0.5 (The test with a variance of 2 is along the x-axis).

$$W = \frac{\frac{SSE_R - SSE_U}{m}}{\frac{SSE_U}{n-k}} = \frac{\frac{SSE_R - SSE_U}{m}}{s^2}. \quad (3.54)$$

where SSE_R is the sum of squared errors of the restricted model.⁹ The restricted model is the original model with the null hypothesis imposed. For example, to test the null $H_0 : \beta_2 = \beta_3 = 0$ against an alternative that $H_1 : \beta_2 \neq 0$ or $\beta_3 \neq 0$ in a bivariate regression,

$$Y_i = \beta_1 + \beta_2 X_{1,i} + \beta_3 X_{2,i} + \varepsilon_i \quad (3.55)$$

⁹The SSE should be the result of minimizing the squared errors. The centered should be used if a constant is included and the uncentered versions if no constant is included.

the restricted model imposes the null,

$$\begin{aligned} Y_i &= \beta_1 + 0X_{1,i} + 0X_{2,i} + \varepsilon_i \\ &= \beta_1 + \varepsilon_i. \end{aligned}$$

The restricted SSE, SSE_R is computed using the residuals from this model while the unrestricted SSE, SSE_U , is computed from the general model that includes both X variables (eq. (3.55)). While Wald tests usually only require the unrestricted model to be estimated, the difference of the SSEs is useful because it can be computed from the output of any standard regression package. Moreover, any linear regression subject to linear restrictions can be estimated using OLS on a modified specification where the constraint is directly imposed. Consider the set of restrictions, \mathbf{R} , in an augmented matrix with \mathbf{r}

$$[\mathbf{R} \quad \mathbf{r}]$$

By transforming this matrix into row-echelon form,

$$[\mathbf{I}_m \quad \tilde{\mathbf{R}} \quad \tilde{\mathbf{r}}]$$

a set of m restrictions can be derived. This also provides a direct method to check whether a set of constraints is logically consistent and feasible or if it contains any redundant restrictions.

Theorem 3.11 (Restriction Consistency and Redundancy). *If $[\mathbf{I}_m \quad \tilde{\mathbf{R}} \quad \tilde{\mathbf{r}}]$ is $[\mathbf{R} \quad \mathbf{r}]$ in reduced echelon form, then a set of restrictions is logically consistent if $\text{rank}(\tilde{\mathbf{R}}) = \text{rank}([\mathbf{I}_m \quad \tilde{\mathbf{R}} \quad \tilde{\mathbf{r}}])$. Additionally, if $\text{rank}(\tilde{\mathbf{R}}) = \text{rank}([\mathbf{I}_m \quad \tilde{\mathbf{R}} \quad \tilde{\mathbf{r}}]) = m$, then there are no redundant restrictions.*

1. Estimate the unrestricted model $Y_i = \mathbf{x}_i\beta + \varepsilon_i$, and the restricted model, $\tilde{Y}_i = \tilde{\mathbf{x}}_i\beta + \varepsilon_i$.
2. Compute $\text{SSE}_R = \sum_{i=1}^n \tilde{\varepsilon}_i^2$ where $\tilde{\varepsilon}_i = \tilde{Y}_i - \tilde{\mathbf{x}}_i\tilde{\beta}$ are the residuals from the restricted regression, and $\text{SSE}_U = \sum_{i=1}^n \hat{\varepsilon}_i^2$ where $\hat{\varepsilon}_i = Y_i - \mathbf{x}_i\hat{\beta}$ are the residuals from the unrestricted regression.
3. Compute $W = \frac{\frac{\text{SSE}_R - \text{SSE}_U}{m}}{\frac{\text{SSE}_U}{n-k}}$.
4. Compare W to the critical value, C_α , of the $F_{m,n-k}$ distribution at size α . Reject the null hypothesis if $W > C_\alpha$.

Finally, in the same sense that the t -stat is a test of the null $H_0 : \beta_k = 0$ against the alternative $H_1 : \beta_k \neq 0$, the F -stat of a regression tests whether all coefficients are zero (except the intercept) against an alternative that at least one is non-zero.

Definition 3.12 (F -stat of a Regression). The F -stat of a regression is the value of a Wald test that all coefficients are zero except the coefficient on the constant (if one is included). Specifically, if the unrestricted model is

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \dots + \beta_k X_{k,i} + \varepsilon_i,$$

the F -stat is the value of a Wald test of the null $H_0 : \beta_2 = \beta_3 = \dots = \beta_k = 0$ against the alternative $H_1 : \beta_j \neq 0$, for $j = 2, \dots, k$ and corresponds to a test based on the restricted regression

$$Y_i = \beta_1 + \varepsilon_i.$$

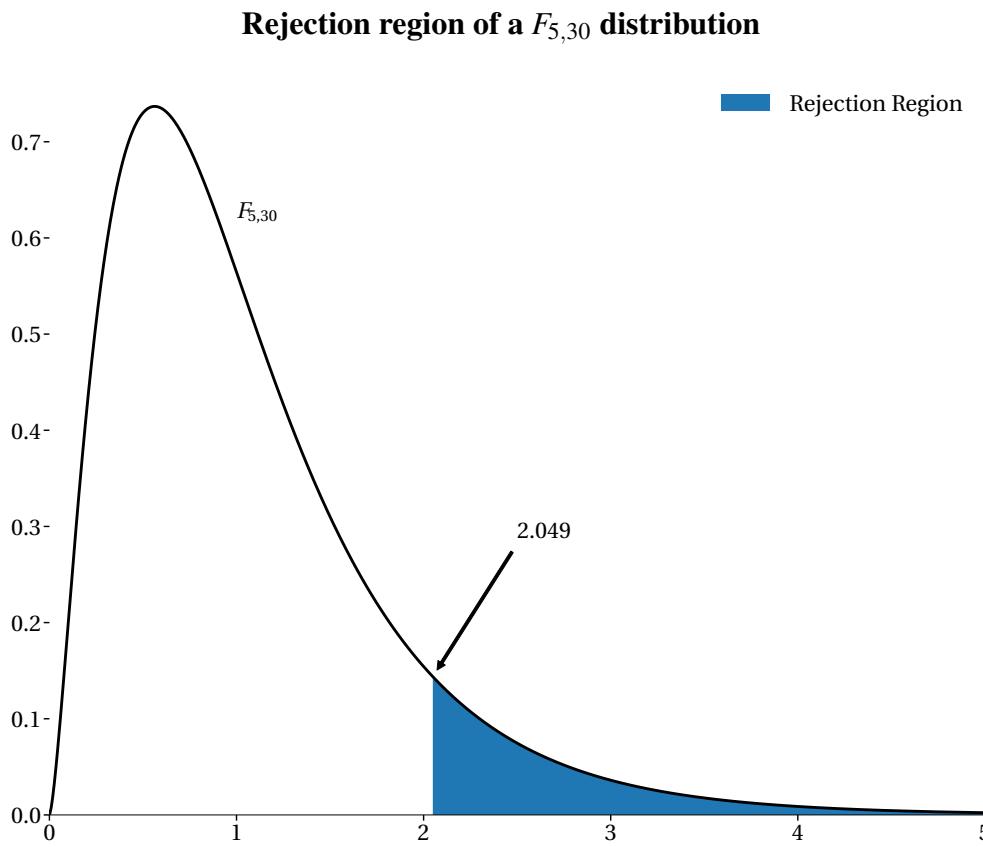


Figure 3.3: Rejection region for a $F_{5,30}$ distribution when using a test with a size of 10%. If the null hypothesis is true, the test statistic should be relatively small (would be 0 if exactly true). Large test statistics lead to rejection of the null hypothesis. In this example, a test statistic with a value greater than 2.049 would lead to a rejection of the null at the 10% level.

3.8.3 Example: T and Wald Tests in Cross-Sectional Factor models

Returning to the factor regression example, the t -stats in the 4-factor model can be computed

$$t_j = \frac{\hat{\beta}_j}{\sqrt{s^2(\mathbf{X}'\mathbf{X})^{-1}_{[jj]}}}.$$

For example, consider a regression of BH^e on the set of four factors and a constant,

$$BH_i^e = \beta_1 + \beta_2 VWM_i^e + \beta_3 SMB_i + \beta_4 HML_i + \beta_5 MOM_i + \varepsilon_i$$

The fit coefficients, t -stats and p-values are contained in table 3.5.

Definition 3.13 (P-value). The p-value is the smallest test size (α) where the null hypothesis may be rejected. The p-value can be equivalently defined as the largest size where the null hypothesis cannot be rejected.

P-values have the advantage that they are independent of the distribution of the test statistic. For example, when using a 2-sided t -test, the p-value of a test statistic t is $2(1 - F_{t_v}(|t|))$ where $F_{t_v}(\cdot)$ is the CDF of a t -distribution with v degrees of freedom. In a Wald test, the p-value is $1 - F_{f_{v_1,v_2}}(W)$ where $F_{f_{v_1,v_2}}(\cdot)$ is the CDF of an f_{v_1,v_2} distribution.

The critical value, C_α , for a 2-sided 10% t -test with 973 degrees of freedom ($n - 5$) is 1.645, and so if $|t| > C_\alpha$ the null hypothesis should be rejected, and the results indicate that the null hypothesis that the coefficients on the constant and SMB are zero cannot be rejected at the 10% level. The p-values indicate the null that the constant was 0 could be rejected at a α of 14% but not one of 13%.

Table 3.5 also contains the Wald test statistics and p-values for a variety of hypotheses, some economically interesting, such as the set of restrictions that the four factor model reduces to the CAPM, $\beta_j = 0$, $j = 1, 3, \dots, 5$. Only one regression, the completely unrestricted regression, was needed to compute all of the test statistics using Wald tests,

$$W = \frac{(\mathbf{R}\beta - \mathbf{r})' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} (\mathbf{R}\beta - \mathbf{r})}{s^2}$$

where \mathbf{R} and \mathbf{r} depend on the null being tested. For example, to test whether a strict CAPM was consistent with the observed data,

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \text{ and } \mathbf{r} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

All of the null hypotheses save one are strongly rejected with p-values of 0 to three decimal places. The sole exception is $H_0 : \beta_1 = \beta_3 = 0$, which produced a Wald test statistic of 2.05. The 5% critical value of an $F_{2,973}$ is 3.005, and so the null hypothesis would be not rejected at the 5% level. The p-value indicates that the test would be rejected at the 13% level but not at the 12% level. One further peculiarity appears in the table. The Wald test statistic for the null $H_0 : \beta_5 = 0$ is exactly the square of the t -test statistic for the same null. This should not be surprising since $W = t^2$ when testing a single linear hypothesis. Moreover, if $z \sim t_v$, then $z^2 \sim F_{1,v}$. This can be seen by inspecting the square of a t_v and applying the definition of an $F_{1,v}$ -distribution.

3.8.4 Likelihood Ratio Tests

Likelihood Ratio (LR) test are based on the relative probability of observing the data if the null is valid to the probability of observing the data under the alternative. The test statistic is defined

$$LR = -2 \ln \left(\frac{\max_{\beta, \sigma^2} f(\mathbf{y} | \mathbf{X}; \beta, \sigma^2)}{\max_{\beta, \sigma^2} f(\mathbf{y} | \mathbf{X}; \beta, \sigma^2)} \right) \quad (3.56)$$

Letting $\hat{\beta}_R$ denote the constrained estimate of β , this test statistic can be reformulated

<i>t</i>-Tests				
	$\hat{\beta}$	s.e. $(\hat{\beta})$	<i>t</i> -stat	<i>p</i> -value
Constant	-0.086	0.042	-2.04	0.042
<i>VWM</i> ^e	1.080	0.010	108.7	0.000
<i>SMB</i>	0.002	0.014	0.13	0.893
<i>HML</i>	0.764	0.015	50.8	0.000
<i>MOM</i>	-0.035	0.010	-3.50	0.000

Wald Tests				
Null	Alternative	<i>W</i>	<i>M</i>	<i>p</i> -value
$\beta_j = 0, j = 2, \dots, 5$	$\beta_j \neq 0, j = 2, \dots, 5$	3558.8	4	0.000
$\beta_j = 0, j = 3, 4, 5$	$\beta_j \neq 0, j = 3, 4, 5$	956.5	3	0.000
$\beta_j = 0, j = 1, 5$	$\beta_j \neq 0, j = 1, 5$	10.1	2	0.000
$\beta_j = 0, j = 1, 3$	$\beta_j \neq 0, j = 1, 3$	2.08	2	0.126
$\beta_5 = 0$	$\beta_5 \neq 0$	12.3	1	0.000

Table 3.5: The upper panel contains *t*-stats and *p*-values for the regression of Big-High excess returns on the four factors and a constant. The lower panel contains test statistics and *p*-values for Wald tests of the reported null hypothesis. Both sets of tests were computed using the small-sample assumptions and may be misleading since the residuals are both non-normal and heteroskedastic.

$$\begin{aligned}
 LR &= -2 \ln \left(\frac{f(\mathbf{y}|\mathbf{X}; \hat{\beta}_R, \hat{\sigma}_R^2)}{f(\mathbf{y}|\mathbf{X}; \hat{\beta}, \hat{\sigma}^2)} \right) \\
 &= -2[l(\hat{\beta}_R, \hat{\sigma}_R^2; \mathbf{y}|\mathbf{X})] - l(\hat{\beta}, \hat{\sigma}^2; \mathbf{y}|\mathbf{X}) \\
 &= 2[l(\hat{\beta}, \hat{\sigma}^2; \mathbf{y}|\mathbf{X}) - l(\hat{\beta}_R, \hat{\sigma}_R^2; \mathbf{y}|\mathbf{X})]
 \end{aligned} \tag{3.57}$$

In the case of the normal log likelihood, *LR* can be further simplified to¹⁰

$$\begin{aligned}
 LR &= -2 \ln \left(\frac{f(\mathbf{y}|\mathbf{X}; \hat{\beta}_R, \hat{\sigma}_R^2)}{f(\mathbf{y}|\mathbf{X}; \hat{\beta}, \hat{\sigma}^2)} \right) \\
 &= -2 \ln \left(\frac{(2\pi\hat{\sigma}_R^2)^{-\frac{n}{2}} \exp(-\frac{(\mathbf{y}-\mathbf{X}\hat{\beta}_R)'(\mathbf{y}-\mathbf{X}\hat{\beta}_R)}{2\hat{\sigma}_R^2})}{(2\pi\hat{\sigma}^2)^{-\frac{n}{2}} \exp(-\frac{(\mathbf{y}-\mathbf{X}\hat{\beta})'(\mathbf{y}-\mathbf{X}\hat{\beta})}{2\hat{\sigma}^2})} \right) \\
 &= -2 \ln \left(\frac{(\hat{\sigma}_R^2)^{-\frac{n}{2}}}{(\hat{\sigma}^2)^{-\frac{n}{2}}} \right) \\
 &= -2 \ln \left(\frac{\hat{\sigma}_R^2}{\hat{\sigma}^2} \right)^{-\frac{n}{2}}
 \end{aligned}$$

¹⁰Note that $\hat{\sigma}_R^2$ and $\hat{\sigma}^2$ use n rather than a degree-of-freedom adjustment since they are MLE estimators.

$$\begin{aligned} &= n [\ln(\hat{\sigma}_R^2) - \ln(\hat{\sigma}^2)] \\ &= n [\ln(\text{SSE}_R) - \ln(\text{SSE}_U)] \end{aligned}$$

Finally, the distribution of the LR statistic can be determined by noting that

$$LR = n \ln \left(\frac{\text{SSE}_R}{\text{SSE}_U} \right) = N \ln \left(\frac{\hat{\sigma}_R^2}{\hat{\sigma}_U^2} \right) \quad (3.58)$$

and that

$$\frac{n-k}{m} \left[\exp \left(\frac{LR}{n} \right) - 1 \right] = W. \quad (3.59)$$

The transformation between W and LR is monotonic so the transformed statistic has the same distribution as W , a $F_{m,n-k}$.

Algorithm 3.2 (Small-Sample Wald Test).

1. Estimate the unrestricted model $Y_i = \mathbf{x}_i\beta + \varepsilon_i$, and the restricted model, $\tilde{Y}_i = \tilde{\mathbf{x}}_i\beta + \varepsilon_i$.
2. Compute $\text{SSE}_R = \sum_{i=1}^n \tilde{\varepsilon}_i^2$ where $\tilde{\varepsilon}_i = \tilde{Y}_i - \tilde{\mathbf{x}}_i\tilde{\beta}$ are the residuals from the restricted regression, and $\text{SSE}_U = \sum_{i=1}^n \hat{\varepsilon}_i^2$ where $\hat{\varepsilon}_i = Y_i - \mathbf{x}_i\hat{\beta}$ are the residuals from the unrestricted regression.
3. Compute $LR = n \ln \left(\frac{\text{SSE}_R}{\text{SSE}_U} \right)$.
4. Compute $W = \frac{n-k}{m} \left[\exp \left(\frac{LR}{n} \right) - 1 \right]$.
5. Compare W to the critical value, C_α , of the $F_{m,n-k}$ distribution at size α . Reject the null hypothesis if $W > C_\alpha$.

3.8.5 Example: LR Tests in Cross-Sectional Factor models

LR tests require estimating the model under both the null and the alternative. In all examples here, the alternative is the unrestricted model with four factors while the restricted models (where the null is imposed) vary. The simplest restricted model corresponds to the most restrictive null, $H_0 : \beta_j = 0$, $j = 1, \dots, 5$, and is specified

$$Y_i = \varepsilon_i.$$

To compute the likelihood ratio, the conditional mean and variance must be estimated. In this simple specification, the conditional mean is $\hat{\mathbf{y}}_R = \mathbf{0}$ (since there are no parameters) and the conditional variance is estimated using the MLE with the mean, $\hat{\sigma}_R^2 = \mathbf{y}'\mathbf{y}/n$ (the sum of squared regressands). The mean under the alternative is $\hat{\mathbf{y}}_U = \mathbf{x}'_i\hat{\beta}$ and the variance is estimated using $\hat{\sigma}_U^2 = (\mathbf{y} - \mathbf{x}'_i\hat{\beta})'(\mathbf{y} - \mathbf{x}'_i\hat{\beta})/n$. Once these quantities have been computed, the LR test statistic is calculated

$$LR = n \ln \left(\frac{\hat{\sigma}_R^2}{\hat{\sigma}_U^2} \right) \quad (3.60)$$

LR Tests					
Null	Alternative	LR	M	p-value	
$\beta_j = 0, j = 2, \dots, 5$	$\beta_j \neq 0, j = 2, \dots, 5$	3558.8	4	0.000	
$\beta_j = 0, j = 3, 4, 5$	$\beta_j \neq 0, j = 3, 4, 5$	956.5	3	0.000	
$\beta_j = 0, j = 1, 5$	$\beta_j \neq 0, j = 1, 5$	10.1	2	0.000	
$\beta_j = 0, j = 1, 3$	$\beta_j \neq 0, j = 1, 3$	2.08	2	0.126	
$\beta_5 = 0$	$\beta_5 \neq 0$	12.3	1	0.000	

LM Tests					
Null	Alternative	LM	M	p-value	
$\beta_j = 0, j = 2, \dots, 5$	$\beta_j \neq 0, j = 2, \dots, 5$	163.4	4	0.000	
$\beta_j = 0, j = 3, 4, 5$	$\beta_j \neq 0, j = 3, 4, 5$	184.3	3	0.000	
$\beta_j = 0, j = 1, 5$	$\beta_j \neq 0, j = 1, 5$	9.85	2	0.000	
$\beta_j = 0, j = 1, 3$	$\beta_j \neq 0, j = 1, 3$	2.07	2	0.127	
$\beta_5 = 0$	$\beta_5 \neq 0$	12.1	1	0.001	

Table 3.6: The upper panel contains test statistics and p-values using LR tests for using a regression of excess returns on the big-high portfolio on the four factors and a constant. In all cases the null was tested against the alternative listed. The lower panel contains test statistics and p-values for LM tests of same tests. Note that the LM test statistics are uniformly smaller than the LR test statistics which reflects that the variance in a LM test is computed from the model estimated under the null, a value that must be larger than the estimate of the variance under the alternative which is used in both the Wald and LR tests. Both sets of tests were computed using the small-sample assumptions and may be misleading since the residuals are non-normal and heteroskedastic.

where the identity $\frac{\hat{\sigma}_R^2}{\hat{\sigma}_U^2} = \frac{SSE_R}{SSE_U}$ has been applied. Finally, LR is transformed by $\frac{n-k}{m} [\exp(\frac{LR}{n}) - 1]$ to produce the test statistic, which is numerically identical to W . This can be seen by comparing the values in table 3.6 to those in table 3.5.

3.8.6 Lagrange Multiplier Tests

Consider minimizing the sum of squared errors subject to a linear hypothesis.

$$\min_{\beta} (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \text{ subject to } \mathbf{R}\beta - \mathbf{r} = 0.$$

This problem can be formulated in terms of a Lagrangian,

$$\mathcal{L}(\beta, \lambda) = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + (\mathbf{R}\beta - \mathbf{r})'\lambda$$

and the problem is

$$\max_{\lambda} \left\{ \min_{\beta} \mathcal{L}(\beta, \lambda) \right\}$$

The first-order conditions correspond to a saddle point,

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \beta} &= -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\beta) + \mathbf{R}'\lambda = \mathbf{0} \\ \frac{\partial \mathcal{L}}{\partial \lambda} &= \mathbf{R}\beta - \mathbf{r} = \mathbf{0}\end{aligned}$$

pre-multiplying the top FOC by $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}$ (which does not change the value, since it is 0),

$$\begin{aligned}2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\beta - 2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} + \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\lambda &= \mathbf{0} \\ \Rightarrow 2\mathbf{R}\beta - 2\mathbf{R}\hat{\beta} + \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\lambda &= \mathbf{0}\end{aligned}$$

where $\hat{\beta}$ is the usual OLS estimator. Solving,

$$\tilde{\lambda} = 2[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta} - \mathbf{r}) \quad (3.61)$$

$$\tilde{\beta} = \hat{\beta} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta} - \mathbf{r}) \quad (3.62)$$

These two solutions provide some insight into the statistical properties of the estimators. $\tilde{\beta}$, the constrained regression estimator, is a function of the OLS estimator, $\hat{\beta}$, and a step in the direction of the constraint. The size of the change is influenced by the distance between the unconstrained estimates and the constraint $(\mathbf{R}\hat{\beta} - \mathbf{r})$. If the unconstrained estimator happened to exactly satisfy the constraint, there would be no step.¹¹

The Lagrange multipliers, $\tilde{\lambda}$, are weighted functions of the unconstrained estimates, $\hat{\beta}$, and will be near zero if the constraint is nearly satisfied ($\mathbf{R}\hat{\beta} - \mathbf{r} \approx 0$). In microeconomics, Lagrange multipliers are known as *shadow prices* since they measure the magnitude of the change in the objective function would if the constraint was relaxed a small amount. Note that $\hat{\beta}$ is the only source of randomness in $\tilde{\lambda}$ (like $\tilde{\beta}$), and so $\tilde{\lambda}$ is a linear combination of normal random variables and will also follow a normal distribution. These two properties combine to provide a mechanism for testing whether the restrictions imposed by the null are consistent with the data. The distribution of $\tilde{\lambda}$ can be directly computed and a test statistic can be formed.

There is another method to derive the LM test statistic that is motivated by the alternative name of LM tests: Score tests. Returning to the first-order conditions and plugging in the parameters,

$$\begin{aligned}\mathbf{R}'\lambda &= 2\mathbf{X}'(\mathbf{y} - \mathbf{X}\tilde{\beta}) \\ \mathbf{R}'\lambda &= 2\mathbf{X}'\tilde{\varepsilon}\end{aligned}$$

where $\tilde{\beta}$ is the constrained estimate of β and $\tilde{\varepsilon}$ are the corresponding estimated errors ($\tilde{\varepsilon} = \mathbf{y} - \mathbf{X}\tilde{\beta}$). Thus $\mathbf{R}'\lambda$ has the same distribution as $2\mathbf{X}'\tilde{\varepsilon}$. However, under the small-sample assumptions, $\tilde{\varepsilon}$ are linear combinations of normal random variables and so are also normal,

$$2\mathbf{X}'\tilde{\varepsilon} \sim N(\mathbf{0}, 4\sigma^2\mathbf{X}'\mathbf{X})$$

¹¹Even if the constraint is valid, the constraint will never be exactly satisfied.

and

$$\mathbf{X}'\tilde{\boldsymbol{\epsilon}} \sim N(\mathbf{0}, \sigma^2 \mathbf{X}'\mathbf{X}). \quad (3.63)$$

A test statistic that the scores are zero can be constructed in the same manner as a Wald test:

$$LM_{\text{Infeasible}} = \frac{\tilde{\boldsymbol{\epsilon}}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\tilde{\boldsymbol{\epsilon}}}{\sigma^2}. \quad (3.64)$$

However, like a Wald test this statistic is not feasible since σ^2 is unknown. Using the same substitution, the LM test statistic is given by

$$LM = \frac{\tilde{\boldsymbol{\epsilon}}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\tilde{\boldsymbol{\epsilon}}}{\tilde{s}^2} \quad (3.65)$$

and has a $F_{m,n-k+m}$ distribution where \tilde{s}^2 is the estimated error variance from the constrained regression. This is a different estimator than was used in constructing a Wald test statistic, where the variance was computed from the unconstrained model. Both estimates are consistent under the null. However, since $SSE_R \geq SSE_U$, \tilde{s}^2 is likely to be larger than s^2 .¹² LM tests are usually implemented using a more convenient – but equivalent – form,

$$LM = \frac{\frac{SSE_R - SSE_U}{m}}{\frac{SSE_R}{n-k+m}}. \quad (3.66)$$

To use the Lagrange Multiplier principle to conduct a test:

Algorithm 3.3 (Small-Sample LM Test).

1. Estimate the unrestricted model $Y_i = \mathbf{x}_i\beta + \varepsilon_i$, and the restricted model, $\tilde{Y}_i = \tilde{\mathbf{x}}_i\beta + \varepsilon_i$.
2. Compute $SSE_R = \sum_{i=1}^n \tilde{\varepsilon}_i^2$ where $\tilde{\varepsilon}_i = \tilde{y}_i - \tilde{\mathbf{x}}_i\tilde{\beta}$ are the residuals from the restricted regression, and $SSE_U = \sum_{i=1}^n \hat{\varepsilon}_i^2$ where $\hat{\varepsilon}_i = Y_i - \mathbf{x}_i\hat{\beta}$ are the residuals from the unrestricted regression.
3. Compute $LM = \frac{\frac{SSE_R - SSE_U}{m}}{\frac{SSE_R}{n-k+m}}$.
4. Compare LM to the critical value, C_α , of the $F_{m,n-k+m}$ distribution at size α . Reject the null hypothesis if $LM > C_\alpha$.

Alternatively, the scores can be directly tested.

Algorithm 3.4 (Alternative Small-Sample LM Test).

1. Estimate the restricted model, $\tilde{Y}_i = \tilde{\mathbf{x}}_i\beta + \varepsilon_i$.
2. Compute $LM = \frac{\frac{\tilde{\boldsymbol{\epsilon}}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\tilde{\boldsymbol{\epsilon}}}{m}}{\frac{\tilde{s}^2}{n-k+m}}$ where \mathbf{X} is n by k the matrix of regressors from the unconstrained model and $s^2 = \frac{\sum_{i=1}^n \tilde{\varepsilon}_i^2}{n-k+m}$.
3. Compare LM to the critical value, C_α , of the $F_{m,n-k+m}$ distribution at size α . Reject the null hypothesis if $LM > C_\alpha$.

¹²Note that since the degree-of-freedom adjustment in the two estimators is different, the magnitude estimated variance is not directly proportional to SSE_R and SSE_U .

3.8.7 Example: LM Tests in Cross-Sectional Factor models

Table 3.6 also contains values from LM tests. LM tests have a slightly different distributions than the Wald and LR and do not produce numerically identical results. While the Wald and LR tests require estimation of the unrestricted model (estimation under the alternative), LM tests only require estimation of the restricted model (estimation under the null). For example, in testing the null $H_0 : \beta_1 = \beta_5 = 0$ (that the *MOM* factor has no explanatory power and that the intercept is 0), the restricted model is estimated from

$$BH_i^e = \gamma_1 VWM_i^e + \gamma_2 SMB_i + \gamma_3 HML_i + \varepsilon_i.$$

The two conditions, that $\beta_1 = 0$ and that $\beta_5 = 0$ are imposed by excluding these regressors. Once the restricted regression is fit, the residuals estimated under the null, $\tilde{\varepsilon}_i = Y_i - \mathbf{x}_i\tilde{\beta}$ are computed and the LM test is calculated from

$$LM = \frac{\tilde{\varepsilon}' \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\tilde{\varepsilon}}{s^2}$$

where \mathbf{X} is the set of explanatory variables from the *unrestricted* regression (in the case, $\mathbf{x}_i = [1 \ VWM_i^e \ SMB_i \ HML_i \ MOM_i]$). Examining table 3.6, the LM test statistics are considerably smaller than the Wald test statistics. This difference arises since the variance used in computing the LM test statistic, $\hat{\sigma}^2$, is estimated under the null. For instance, in the most restricted case ($H_0 = \beta_j = 0$, $j = 1, \dots, k$), the variance is estimated by $\mathbf{y}'\mathbf{y}/N$ (since $k = 0$ in this model) which is very different from the variance estimated under the alternative (which is used by both the Wald and LR). Despite the differences in the test statistics, the p-values in the table would result in the same inference. For the one hypothesis that is not completely rejected, the p-value of the LM test is slightly larger than that of the LR (or W). However, .130 and .129 should never make a qualitative difference (nor should .101 and .099, even when using a 10% test). These results highlight a general feature of LM tests: test statistics based on the LM-principle are smaller than Likelihood Ratios and Wald tests, and so less likely to reject.

3.8.8 Comparing the Wald, LR, and LM Tests

With three tests available to test the same hypothesis, which is the correct one? In the small-sample framework, the Wald is the obvious choice because $W \approx LR$ and W is larger than LM . However, the LM has a slightly different distribution, so it is impossible to make an absolute statement. The choice among these three tests reduces to user preference and ease of computation. Since computing SSE_U and SSE_R is simple, the Wald test is likely the simplest to implement.

These results are no longer true when nonlinear restrictions and/or nonlinear models are estimated. Further discussion of the factors affecting the choice between the Wald, LR, and LM tests will be reserved until then. Figure 3.4 contains a graphical representation of the three test statistics in the context of a simple regression, $Y_i = \beta X_i + \varepsilon_i$.¹³ The Wald test measures the magnitude of the constraint $R\beta - r$ at the unconstrained estimator $\hat{\beta}$. The LR test measures how much of the sum of squared errors has changed between $\hat{\beta}$ and $\tilde{\beta}$. Finally, the LM test measures the magnitude of the gradient, $\mathbf{X}'(\mathbf{y} - \mathbf{X}\beta)$ at the constrained estimator $\tilde{\beta}$.

¹³Magnitudes of the lines is not to scale, so the magnitude of the test statistics cannot be determined from the picture.

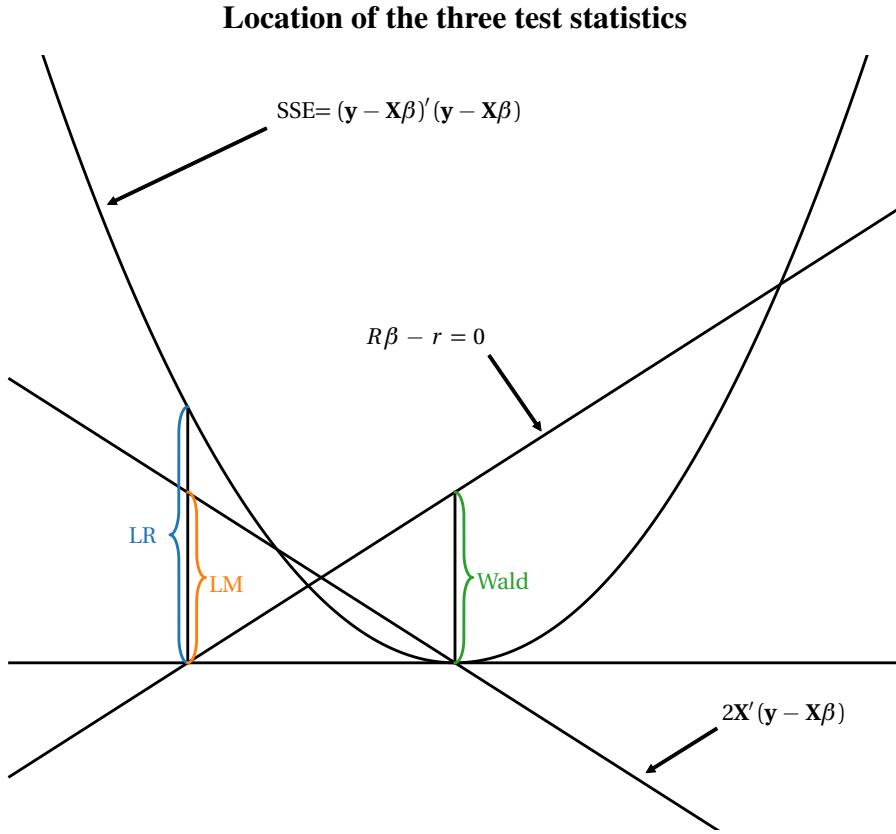


Figure 3.4: Graphical representation of the three major classes of tests. The Wald test measures the magnitude of the constraint, $\mathbf{R}\beta - r$, at the OLS parameter estimate, $\hat{\beta}$. The LM test measures the magnitude of the score at the restricted estimator ($\tilde{\beta}$) while the LR test measures the difference between the SSE at the restricted estimator and the SSE at the unrestricted estimator. Note: Only the location of the test statistic, not their relative magnitudes, can be determined from this illustration.

3.9 Large-Sample Assumption

While the small-sample assumptions allow the exact distribution of the OLS estimator and test statistics to be derived, these assumptions are not realistic in applications using financial data. Asset returns are non-normal (both skewed and leptokurtic), heteroskedastic, and correlated. The large-sample framework allows for inference on β without making strong assumptions about the distribution or error covariance structure. However, the generality of the large-sample framework comes at the loss of the ability to say anything exact about the estimates in finite samples.

Four new assumptions are needed to analyze the asymptotic behavior of the OLS estimators.

Assumption 3.7 (Stationary Ergodicity). $\{(\mathbf{x}_i, \varepsilon_i)\}$ is a strictly stationary and ergodic sequence.

This is a technical assumption needed for consistency and asymptotic normality. It implies two properties about the joint density of $\{(\mathbf{x}_i, \varepsilon_i)\}$: the joint distribution of $\{(\mathbf{x}_i, \varepsilon_i)\}$ and $\{(\mathbf{x}_{i+j}, \varepsilon_{i+j})\}$ depends on the time between observations (j) and *not* the observation index (i) and that averages will converge to their expected value (as long as they exist). There are a number of alternative assumptions

that could be used in place of this assumption, although this assumption is broad enough to allow for i.i.d., i.d.n.d (independent not identically distributed, including heteroskedasticity), and some n.i.n.i.d. data, although it does rule out some important cases. Specifically, the regressors cannot be trending or otherwise depend on the observation index, an important property of some economic time series such as the level of a market index or aggregate consumption. Stationarity will be considered more carefully in the time-series chapters.

Assumption 3.8 (Rank). $E[\mathbf{x}_i' \mathbf{x}_i] = \Sigma_{\mathbf{XX}}$ is nonsingular and finite.

This assumption, like assumption 3.3, is needed to ensure identification.

Assumption 3.9 (Martingale Difference). $\{\mathbf{x}_i' \boldsymbol{\varepsilon}_i, \mathcal{F}_i\}$ is a martingale difference sequence,

$$E \left[(X_{j,i} \boldsymbol{\varepsilon}_i)^2 \right] < \infty, \quad j = 1, 2, \dots, k, \quad i = 1, 2, \dots$$

and $\mathbf{S} = V[n^{-1/2} \mathbf{X}' \boldsymbol{\varepsilon}]$ is finite and non singular.

A martingale difference sequence has the property that its mean is unpredictable using the information contained in the information set (\mathcal{F}_i).

Definition 3.14 (Martingale Difference Sequence). Let $\{\mathbf{Z}_i\}$ be a vector stochastic process and \mathcal{F}_i be the information set corresponding to observation i containing all information available when observation i was collected except \mathbf{Z}_i . $\{\mathbf{Z}_i, \mathcal{F}_i\}$ is a martingale difference sequence if

$$E[\mathbf{Z}_i | \mathcal{F}_i] = \mathbf{0}$$

In the context of the linear regression model, it states that the current score is not predictable by any of the previous scores, that the mean of the scores is zero ($E[\mathbf{X}_i' \boldsymbol{\varepsilon}_i] = 0$), and there is no other variable in \mathcal{F}_i which can predict the scores. This assumption is sufficient to ensure that $n^{-1/2} \mathbf{X}' \boldsymbol{\varepsilon}$ will follow a Central Limit Theorem, and it plays a role in consistency of the estimator. A m.d.s. is a fairly general construct and does not exclude using time-series regressors as long as they are predetermined, meaning that they do not depend on the process generating $\boldsymbol{\varepsilon}_i$. For instance, in the CAPM, the return on the market portfolio can be thought of as being determined independently of the idiosyncratic shock affecting individual assets.

Assumption 3.10 (Moment Existence). $E[X_{j,i}^4] < \infty, \quad i = 1, 2, \dots, \quad j = 1, 2, \dots, k$ and $E[\boldsymbol{\varepsilon}_i^2] = \sigma^2 < \infty, \quad i = 1, 2, \dots$

This final assumption requires that the fourth moment of any regressor exists and the variance of the errors is finite. This assumption is needed to derive a consistent estimator of the parameter covariance.

3.10 Large-Sample Properties

These assumptions lead to two theorems that describe the asymptotic behavior of $\hat{\beta}$: it is consistent and asymptotically normally distributed. First, some new notation is needed. Let

$$\hat{\beta}_n = \left(\frac{\mathbf{X}' \mathbf{X}}{n} \right)^{-1} \left(\frac{\mathbf{X}' \mathbf{y}}{n} \right) \quad (3.67)$$

be the regression coefficient using n realizations from the stochastic process $\{\mathbf{x}_i, \boldsymbol{\varepsilon}_i\}$.

Theorem 3.12 (Consistency of $\hat{\beta}$). *Under assumptions 3.1 and 3.7 - 3.9*

$$\hat{\beta}_n \xrightarrow{p} \beta$$

Consistency is a weak property of the OLS estimator, but it is important. This result relies crucially on the implication of assumption 3.9 that $n^{-1}\mathbf{X}'\varepsilon \xrightarrow{p} \mathbf{0}$, and under the same assumptions, the OLS estimator is also asymptotically normally distributed.

Theorem 3.13 (Asymptotic Normality of $\hat{\beta}$). *Under assumptions 3.1 and 3.7 - 3.9*

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \Sigma_{\mathbf{XX}}^{-1} \mathbf{S} \Sigma_{\mathbf{XX}}^{-1}) \quad (3.68)$$

where $\Sigma_{\mathbf{XX}} = E[\mathbf{x}_i' \mathbf{x}_i]$ and $\mathbf{S} = V[n^{-1/2} \mathbf{X}' \varepsilon]$

Asymptotic normality provides the basis for hypothesis tests on β . However, using only theorem 3.13, tests are not feasible since $\Sigma_{\mathbf{XX}}$ and \mathbf{S} are unknown, and so must be estimated.

Theorem 3.14 (Consistency of OLS Parameter Covariance Estimator). *Under assumptions 3.1 and 3.7 - 3.10,*

$$\begin{aligned} \hat{\Sigma}_{\mathbf{XX}} &= n^{-1} \mathbf{X}' \mathbf{X} \xrightarrow{p} \Sigma_{\mathbf{XX}} \\ \hat{\mathbf{S}} &= n^{-1} \sum_{i=1}^n e_i^2 \mathbf{x}_i' \mathbf{x}_i \xrightarrow{p} \mathbf{S} \\ &= n^{-1} (\mathbf{X}' \hat{\mathbf{E}} \mathbf{X}) \end{aligned}$$

and

$$\hat{\Sigma}_{\mathbf{XX}}^{-1} \hat{\mathbf{S}} \hat{\Sigma}_{\mathbf{XX}}^{-1} \xrightarrow{p} \Sigma_{\mathbf{XX}}^{-1} \mathbf{S} \Sigma_{\mathbf{XX}}^{-1}$$

where $\hat{\mathbf{E}} = \text{diag}(\hat{\varepsilon}_1^2, \dots, \hat{\varepsilon}_n^2)$ is a matrix with the estimated residuals squared along its diagonal.

Combining these theorems, the OLS estimator is consistent, asymptotically normal, and the asymptotic variance can be consistently estimated. These three properties provide the tools necessary to conduct hypothesis tests in the asymptotic framework. The usual estimator of the residual variance is also consistent for the variance of the innovations under the same conditions.

Theorem 3.15 (Consistency of OLS Variance Estimator). *Under assumptions 3.1 and 3.7 - 3.10 ,*

$$\hat{\sigma}_n^2 = n^{-1} \hat{\varepsilon}' \hat{\varepsilon} \xrightarrow{p} \sigma^2$$

Further, if homoskedasticity is assumed, then the parameter covariance estimator can be simplified.

Theorem 3.16 (Homoskedastic Errors). *Under assumptions 3.1, 3.4, 3.5 and 3.7 - 3.10,*

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \sigma^2 \Sigma_{\mathbf{XX}}^{-1})$$

Combining the result of this theorem with that of theorems 3.14 and 3.15, a consistent estimator of $\sigma^2 \Sigma_{\mathbf{XX}}^{-1}$ is given by $\hat{\sigma}_n^2 \hat{\Sigma}_{\mathbf{XX}}^{-1}$.

3.11 Large-Sample Hypothesis Testing

All three test types, Wald, LR, and LM, have large-sample equivalents that exploit the estimated parameters' asymptotic normality. While these tests are only asymptotically exact, the use of the asymptotic distribution is justified as an approximation to the finite-sample distribution, although the quality of the CLT approximation depends on how well behaved the data are.

3.11.1 Wald Tests

Recall from Theorem 3.13,

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \Sigma_{\mathbf{xx}}^{-1} \mathbf{S} \Sigma_{\mathbf{xx}}^{-1}). \quad (3.69)$$

Applying the properties of a normal random variable, if $\mathbf{z} \sim N(\mu, \Sigma)$, $\mathbf{c}'\mathbf{z} \sim N(\mathbf{c}'\mu, \mathbf{c}'\Sigma\mathbf{c})$ and that if $w \sim N(\mu, \sigma^2)$ then $\frac{(w-\mu)^2}{\sigma^2} \sim \chi_1^2$. Using these two properties, a test of the null

$$H_0 : \mathbf{R}\beta - \mathbf{r} = 0$$

against the alternative

$$H_1 : \mathbf{R}\beta - \mathbf{r} \neq 0$$

can be constructed.

Following from Theorem 3.13, if $H_0 : \mathbf{R}\beta - \mathbf{r} = 0$ is true, then

$$\sqrt{n}(\mathbf{R}\hat{\beta}_n - \mathbf{r}) \xrightarrow{d} N(0, \mathbf{R}\Sigma_{\mathbf{xx}}^{-1} \mathbf{S} \Sigma_{\mathbf{xx}}^{-1} \mathbf{R}') \quad (3.70)$$

and

$$\Gamma^{-\frac{1}{2}} \sqrt{n}(\mathbf{R}\hat{\beta}_n - \mathbf{r}) \xrightarrow{d} N(0, \mathbf{I}_k) \quad (3.71)$$

where $\Gamma = \mathbf{R}\Sigma_{\mathbf{xx}}^{-1} \mathbf{S} \Sigma_{\mathbf{xx}}^{-1} \mathbf{R}'$. Under the null that $H_0 : \mathbf{R}\beta - \mathbf{r} = 0$,

$$n(\mathbf{R}\hat{\beta}_n - \mathbf{r})' [\mathbf{R}\Sigma_{\mathbf{xx}}^{-1} \mathbf{S} \Sigma_{\mathbf{xx}}^{-1} \mathbf{R}']^{-1} (\mathbf{R}\hat{\beta}_n - \mathbf{r}) \xrightarrow{d} \chi_m^2 \quad (3.72)$$

where m is the rank(\mathbf{R}). This estimator is not feasible since Γ is not known and must be estimated. Fortunately, Γ can be consistently estimated by applying the results of Theorem 3.14

$$\hat{\Sigma}_{\mathbf{xx}} = n^{-1} \mathbf{X}' \mathbf{X}$$

$$\hat{\mathbf{S}} = n^{-1} \sum_{i=1}^n e_i^2 \mathbf{x}_i' \mathbf{x}_i$$

and so

$$\hat{\Gamma} = \hat{\Sigma}_{\mathbf{xx}}^{-1} \hat{\mathbf{S}} \hat{\Sigma}_{\mathbf{xx}}^{-1}.$$

The feasible Wald statistic is defined

$$W = n(\mathbf{R}\hat{\beta}_n - \mathbf{r})' \left[\mathbf{R}\hat{\Sigma}_{\mathbf{XX}}^{-1} \hat{\mathbf{S}}\hat{\Sigma}_{\mathbf{XX}}^{-1} \mathbf{R}' \right]^{-1} (\mathbf{R}\hat{\beta}_n - \mathbf{r}) \xrightarrow{d} \chi_m^2. \quad (3.73)$$

Test statistic values can be compared to the critical value C_α from a χ_m^2 at the α -significance level and the null is rejected if W is greater than C_α . The asymptotic t -test (which has a normal distribution) is defined analogously,

$$t = \sqrt{n} \frac{\mathbf{R}\hat{\beta}_n - r}{\sqrt{\mathbf{R}\hat{\Gamma}\mathbf{R}'}} \xrightarrow{d} N(0, 1), \quad (3.74)$$

where \mathbf{R} is a 1 by k vector. Typically \mathbf{R} is a vector with 1 in its j^{th} element, producing statistic

$$t = \sqrt{n} \frac{\hat{\beta}_{jN}}{\sqrt{[\hat{\Gamma}]_{jj}}} \xrightarrow{d} N(0, 1)$$

where $[\hat{\Gamma}]_{jj}$ is the j^{th} diagonal element of $\hat{\Gamma}$.

The n term in the Wald statistic (or \sqrt{n} in the t -test) may appear strange at first, although these terms are also present in the classical tests. Recall that the t -stat (null $H_0 : \beta_j = 0$) from the classical framework with homoskedastic data is given by

$$t_1 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}}}.$$

The t -stat in the asymptotic framework is

$$t_2 = \sqrt{n} \frac{\hat{\beta}_{jN}}{\sqrt{\hat{\sigma}^2[\hat{\Sigma}_{\mathbf{XX}}^{-1}]_{jj}}}.$$

If t_1 is multiplied and divided by \sqrt{n} , then

$$t_1 = \sqrt{n} \frac{\hat{\beta}_j}{\sqrt{n} \sqrt{\hat{\sigma}^2[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}}} = \sqrt{n} \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2[(\frac{\mathbf{X}'\mathbf{X}}{n})^{-1}]_{jj}}} = \sqrt{n} \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2[\hat{\Sigma}_{\mathbf{XX}}^{-1}]_{jj}}} = t_2,$$

and these two statistics have the same value since $\mathbf{X}'\mathbf{X}$ differs from $\hat{\Sigma}_{\mathbf{XX}}$ by a factor of n .

Algorithm 3.5 (Large-Sample Wald Test).

1. Estimate the unrestricted model $Y_i = \mathbf{X}_i\beta + \varepsilon_i$.
2. Estimate the parameter covariance using $\hat{\Sigma}_{\mathbf{XX}}^{-1} \hat{\mathbf{S}}\hat{\Sigma}_{\mathbf{XX}}^{-1}$ where

$$\hat{\Sigma}_{\mathbf{XX}} = n^{-1} \sum_{i=1}^n \mathbf{x}'_i \mathbf{x}_i, \quad \hat{\mathbf{S}} = n^{-1} \sum_{i=1}^n \hat{\varepsilon}_i^2 \mathbf{x}'_i \mathbf{x}_i$$

3. Construct the restriction matrix, \mathbf{R} , and the value of the restriction, \mathbf{r} , from the null hypothesis.
4. Compute $W = n(\mathbf{R}\hat{\beta}_n - \mathbf{r})' \left[\mathbf{R}\hat{\Sigma}_{\mathbf{XX}}^{-1} \hat{\mathbf{S}}\hat{\Sigma}_{\mathbf{XX}}^{-1} \mathbf{R}' \right]^{-1} (\mathbf{R}\hat{\beta}_n - \mathbf{r})$.
5. Reject the null if $W > C_\alpha$ where C_α is the critical value from a χ_m^2 using a size of α .

3.11.2 Lagrange Multiplier Tests

Recall that the first-order conditions of the constrained estimation problem require

$$\mathbf{R}'\hat{\lambda} = 2\mathbf{X}'\tilde{\varepsilon}$$

where $\tilde{\varepsilon}$ are the residuals estimated under the null $H_0 : \mathbf{R}\beta - \mathbf{r} = 0$. The LM test examines whether $\hat{\lambda}$ is close to zero. In the large-sample framework, $\hat{\lambda}$, like $\hat{\beta}$, is asymptotically normal and $\mathbf{R}'\hat{\lambda}$ will only be close to 0 if $\hat{\lambda} \approx 0$. The asymptotic version of the LM test can be compactly expressed if $\tilde{\mathbf{s}}$ is defined as the average score of the restricted estimator, $\tilde{\mathbf{s}} = n^{-1}\mathbf{X}'\tilde{\varepsilon}$. In this notation,

$$LM = n\tilde{\mathbf{s}}'\mathbf{S}^{-1}\tilde{\mathbf{s}} \xrightarrow{d} \chi_m^2. \quad (3.75)$$

If the model is correctly specified, $n^{-1}\mathbf{X}'\tilde{\varepsilon}$, which is a k by 1 vector with j^{th} element $n^{-1} \sum_{i=1}^n x_{j,i}\tilde{\varepsilon}_i$, should be a mean-zero vector with asymptotic variance \mathbf{S} by assumption 3.7. Thus, $\sqrt{n}(n^{-1}\mathbf{X}'\tilde{\varepsilon}) \xrightarrow{d} N(0, \mathbf{S})$ implies

$$\sqrt{n}\mathbf{S}^{-\frac{1}{2}}\tilde{\mathbf{s}} \xrightarrow{d} N\left(\mathbf{0}, \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\right) \quad (3.76)$$

and so $n\tilde{\mathbf{s}}'\mathbf{S}^{-1}\tilde{\mathbf{s}} \xrightarrow{d} \chi_m^2$. This version is infeasible and the feasible version of the LM test must be used,

$$LM = n\tilde{\mathbf{s}}'\tilde{\mathbf{S}}^{-1}\tilde{\mathbf{s}} \xrightarrow{d} \chi_m^2. \quad (3.77)$$

where $\tilde{\mathbf{S}} = n^{-1} \sum_{i=1}^n \tilde{\varepsilon}_i^2 \mathbf{x}_i' \mathbf{x}_i$ is the estimator of the asymptotic variance *computed under the null*. This means that $\tilde{\mathbf{S}}$ is computed using the residuals from the restricted regression, $\tilde{\varepsilon}$, and that it will differ from the usual estimator $\hat{\mathbf{S}}$ which is computed using residuals from the unrestricted regression, $\hat{\varepsilon}$. Under the null, both $\tilde{\mathbf{S}}$ and $\hat{\mathbf{S}}$ are consistent estimators for \mathbf{S} and using one or the other has no asymptotic effect.

If the residuals are homoskedastic, the LM test can also be expressed in terms of the R^2 of the unrestricted model when testing a null that the coefficients on all explanatory variables except the intercept are zero. Suppose the regression fit was

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \dots + \beta_k X_{kn}.$$

To test the $H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$ (where the excluded β_1 corresponds to a constant),

$$LM = nR^2 \xrightarrow{d} \chi_k^2 \quad (3.78)$$

is equivalent to the test statistic in eq. (3.77). This expression is useful as a simple tool to test whether the explanatory variables in a regression appear to explain *any* variation in the dependent variable. If the residuals are heteroskedastic, the nR^2 form of the LM test does not have standard distribution and should not be used.

Algorithm 3.6 (Large-Sample LM Test).

1. Form the unrestricted model, $Y_i = \mathbf{X}_i\beta + \varepsilon_i$.
2. Impose the null on the unrestricted model and estimate the restricted model, $\tilde{Y}_i = \tilde{\mathbf{X}}_i\beta + \varepsilon_i$.

3. Compute the residuals from the restricted regression, $\tilde{\epsilon}_i = \tilde{Y}_i - \tilde{\mathbf{x}}_i \tilde{\beta}$.
4. Construct the score using the residuals from the restricted regression from both models, $\tilde{\mathbf{s}}_i = \mathbf{x}_i \tilde{\epsilon}_i$ where \mathbf{x}_i are the regressors from the unrestricted model.
5. Estimate the average score and the covariance of the score,

$$\tilde{\mathbf{s}} = n^{-1} \sum_{i=1}^n \tilde{\mathbf{s}}_i, \quad \tilde{\mathbf{S}} = n^{-1} \sum_{i=1}^n \tilde{\mathbf{s}}_i' \tilde{\mathbf{s}}_i \quad (3.79)$$

6. Compute the LM test statistic as $LM = n\tilde{\mathbf{s}}'\tilde{\mathbf{S}}^{-1}\tilde{\mathbf{s}}$.
7. Reject the null if $LM > C_\alpha$ where C_α is the critical value from a χ_m^2 using a size of α .

3.11.3 Likelihood Ratio Tests

A critical distinction between small-sample and large-sample hypothesis testing is the omission of assumption 3.6. Without this assumption, the distribution of the errors is left unspecified. Based on the ease of implementing the Wald and LM tests their asymptotic framework, it may be tempting to think the likelihood ratio is asymptotically valid. It is not. The technical details are complicated, and the validity of the asymptotic distribution of the LR relies crucially on the Information Matrix Equality holding. If the shocks are heteroskedastic, then the IME will generally not hold, and the distribution of LR tests will be nonstandard.¹⁴

There is, however, a feasible likelihood-ratio like test available. The motivation for this test will be clarified in the GMM chapter. For now, the functional form will be given with only minimal explanation,

$$LR = n\tilde{\mathbf{s}}'\tilde{\mathbf{S}}^{-1}\tilde{\mathbf{s}} \xrightarrow{d} \chi_m^2, \quad (3.80)$$

where $\tilde{\mathbf{s}} = n^{-1}\mathbf{X}'\tilde{\epsilon}$ is the average score vector when the estimator is computed under the null. This statistic is similar to the LM test statistic, although there are two differences. First, one term has been left out of this expression, and the formal definition of the asymptotic LR is

$$LR = n\tilde{\mathbf{s}}'\tilde{\mathbf{S}}^{-1}\tilde{\mathbf{s}} - \hat{\mathbf{s}}'\hat{\mathbf{S}}^{-1}\hat{\mathbf{s}} \xrightarrow{d} \chi_m^2 \quad (3.81)$$

where $\hat{\mathbf{s}} = n^{-1}\mathbf{X}'\hat{\epsilon}$ are the average scores from the *unrestricted* estimator. Recall from the first-order conditions of OLS (eq. (3.7)) that $\hat{\mathbf{s}} = \mathbf{0}$ and the second term in the general expression of the *LR* will always be zero. The second difference between *LR* and *LM* exists only in the feasible versions. The feasible version of the LR is given by

$$LR = n\tilde{\mathbf{s}}'\hat{\mathbf{S}}^{-1}\tilde{\mathbf{s}} \xrightarrow{d} \chi_m^2. \quad (3.82)$$

where $\hat{\mathbf{S}}$ is estimated using the scores of the *unrestricted* model (under the alternative),

$$\hat{\mathbf{S}}^{-1} = \frac{1}{n} \sum_{i=1}^n \hat{\epsilon}_i^2 \mathbf{x}_i' \mathbf{x}_i. \quad (3.83)$$

¹⁴In this case, the LR will converge to a weighted mixture of m independent χ_1^2 random variables where the weights are not 1. The resulting distribution is not a χ_m^2 .

The feasible LM, $n\tilde{\mathbf{s}}'\hat{\mathbf{S}}^{-1}\tilde{\mathbf{s}}$, uses a covariance estimator ($\hat{\mathbf{S}}$) based on the scores from the restricted model, $\tilde{\mathbf{s}}$.

In models with heteroskedasticity, it is impossible to determine *a priori* whether the LM or the LR test statistic will be larger, although folk wisdom states that LR test statistics are larger than LM test statistics (and hence the LR will be more powerful). If the data are homoskedastic, and homoskedastic estimators of $\hat{\mathbf{S}}$ and $\tilde{\mathbf{S}}$ are used ($\hat{\sigma}^2(\mathbf{X}'\mathbf{X}/n)^{-1}$ and $\tilde{\sigma}^2(\mathbf{X}'\mathbf{X}/n)^{-1}$, respectively), then it must be the case that $LM < LR$. This ordering of the two test statistic occurs since $\hat{\sigma}^2$ must be smaller than $\tilde{\sigma}^2$ because OLS minimizes the squared residuals. The LR is guaranteed to have more power in this case.

Algorithm 3.7 (Large-Sample LR Test).

1. Estimate the unrestricted model $Y_i = \mathbf{X}_i\beta + \varepsilon_i$.
2. Impose the null on the unrestricted model and estimate the restricted model, $\tilde{Y}_i = \tilde{\mathbf{X}}_i\beta + \varepsilon_i$.
3. Compute the residuals from the restricted regression, $\tilde{\varepsilon}_i = \tilde{Y}_i - \tilde{\mathbf{x}}_i\tilde{\beta}$, and from the unrestricted regression, $\hat{\varepsilon}_i = Y_i - \mathbf{x}_i\hat{\beta}$.
4. Construct the score from both models, $\tilde{\mathbf{s}}_i = \mathbf{x}_i\tilde{\varepsilon}_i$ and $\hat{\mathbf{s}}_i = \mathbf{x}_i\hat{\varepsilon}_i$, where in both cases \mathbf{x}_i are the regressors from the unrestricted model.
5. Estimate the average score and the covariance of the score,

$$\tilde{\mathbf{s}} = n^{-1} \sum_{i=1}^n \tilde{\mathbf{s}}_i, \quad \hat{\mathbf{S}} = n^{-1} \sum_{i=1}^n \hat{\mathbf{s}}_i' \hat{\mathbf{s}}_i \quad (3.84)$$

6. Compute the LR test statistic as $LR = n\tilde{\mathbf{s}}'\hat{\mathbf{S}}^{-1}\tilde{\mathbf{s}}$.
7. Reject the null if $LR > C_\alpha$ where C_α is the critical value from a χ_m^2 using a size of α .

3.11.4 Revisiting the Wald, LM, and LR tests

The previous tests can now be revisited while allowing for heteroskedasticity in the data. Tables 3.7 and 3.8 contain t -tests, Wald tests, LM tests, and LR tests that compare large-sample versions of these test statistics to their small-sample framework equivalents. There is a clear direction in the difference between the small-sample and large-sample test statistics: the large-sample statistics are smaller than the small-sample statistics, often substantially. Examining table 3.7, 4 out of 5 of the t -stats have decreased. Since the estimator of $\hat{\beta}$ is the same in both the small-sample and the large-sample frameworks, all of the difference is attributable to changes in the standard errors, which typically increased by 50%. When t -stats differ dramatically under the two covariance estimators, the likely cause is heteroskedasticity.

Table 3.8 shows that the Wald, LR, and LM test statistics also changed by large amounts.¹⁵ The heteroskedasticity-robust Wald statistics decreased by up to a factor of 2, and the robust LM test statistics decreased by up to 5 times. The LR test statistic values were generally larger than those

¹⁵The statistics based on the small-sample assumptions have $f_{m,t-k}$ or $f_{m,t-k+m}$ distributions while the statistics based on the large-sample assumptions have χ_m^2 distributions, and so the values of the small-sample statistics must be multiplied by m to be compared to the large-sample statistics.

	$\hat{\beta}$	Homoskedasticity			Heteroskedasticity		
		s.e.($\hat{\beta}$)	t-stat	p-value	s.e.($\hat{\beta}$)	t-stat	p-value
Constant	-0.086	0.042	-2.04	0.042	0.043	-1.991	0.046
VWM^e	1.080	0.010	108.7	0.000	0.012	93.514	0.000
SMB	0.002	0.014	0.13	0.893	0.017	0.110	0.912
HML	0.764	0.015	50.8	0.000	0.021	36.380	0.000
MOM	-0.035	0.010	-3.50	0.000	0.013	-2.631	0.009

Table 3.7: Comparing small and large-sample t -stats. The small-sample statistics in the left panel of the table overstate the precision of the estimates. The heteroskedasticity robust standard errors are larger for 4 out of 5 parameters, and one variable which was significant at the 15% level is insignificant.

of the corresponding Wald or LR test statistics. The relationship between the robust versions of the Wald and LR statistics is not clear, and for models that are grossly misspecified, the Wald and LR test statistics are substantially larger than their LM counterparts. However, when the value of the test statistics is smaller, the three are virtually identical, and the decision taken using any of these three tests is the same. All nulls except $H_0 : \beta_1 = \beta_3 = 0$ are rejected using standard sizes (5-10%).

These changes should serve as a warning to conducting inference using covariance estimates based on homoskedasticity. In most applications to financial time-series, heteroskedasticity robust covariance estimators (and often HAC (Heteroskedasticity and Autocorrelation Consistent), which will be defined in the time-series chapter) are automatically applied without testing for heteroskedasticity.

3.12 Violations of the Large-Sample Assumptions

The large-sample assumptions are just that: assumptions. While this set of assumptions is far more general than the finite-sample setup, they may be violated in a number of ways. This section examines the consequences of certain violations of the large-sample assumptions.

3.12.1 Omitted and Extraneous Variables

Suppose that the model is linear but misspecified, and a subset of the relevant regressors are excluded. The model can be specified

$$Y_i = \beta_1 \mathbf{X}_{1,i} + \beta_2 \mathbf{X}_{2,i} + \varepsilon_i \quad (3.85)$$

where $\mathbf{X}_{1,i}$ is 1 by k_1 vector of included regressors and $\mathbf{X}_{2,i}$ is a 1 by k_2 vector of excluded but relevant regressors. Omitting $\mathbf{x}_{2,i}$ from the fit model, the least-squares estimator is

$$\hat{\beta}_{1n} = \left(\frac{\mathbf{X}'_1 \mathbf{X}_1}{n} \right)^{-1} \frac{\mathbf{X}'_1 \mathbf{y}}{n}. \quad (3.86)$$

This misspecified estimator is biased, and the bias depends on the magnitude of the coefficients on the omitted variables and the correlation between the omitted and excluded regressors.

		Wald Tests			LR Tests		
Null	Alternative	M	Small Sample		Large Sample		
			W	p-value	W	p-value	
$\beta_j = 0, j = 2, \dots, 5$	$\beta_j \neq 0, j = 2, \dots, 5$	4	3558.8	0.000	2661.2	0.000	
$\beta_j = 0, j = 3, 4, 5$	$\beta_j \neq 0, j = 3, 4, 5$	3	956.5	0.000	583.2	0.000	
$\beta_j = 0, j = 1, 5$	$\beta_j \neq 0, j = 1, 5$	2	10.1	0.000	7.35	0.001	
$\beta_j = 0, j = 1, 3$	$\beta_j \neq 0, j = 1, 3$	2	2.08	0.126	2.04	0.131	
$\beta_5 = 0$	$\beta_5 \neq 0$	1	12.3	0.000	6.92	0.009	

		LM Tests			Large Sample		
Null	Alternative	M	Small Sample		Large Sample		
			LM	p-value	LM	p-value	
$\beta_j = 0, j = 2, \dots, 5$	$\beta_j \neq 0, j = 2, \dots, 5$	4	163.4	0.000	34.8	0.000	
$\beta_j = 0, j = 3, 4, 5$	$\beta_j \neq 0, j = 3, 4, 5$	3	184.3	0.000	31.9	0.000	
$\beta_j = 0, j = 1, 5$	$\beta_j \neq 0, j = 1, 5$	2	9.85	0.000	7.82	0.000	
$\beta_j = 0, j = 1, 3$	$\beta_j \neq 0, j = 1, 3$	2	2.07	0.127	2.11	0.121	
$\beta_5 = 0$	$\beta_5 \neq 0$	1	12.1	0.001	6.50	0.011	

Table 3.8: Comparing large- and small-sample Wald, LM, and LR test statistics. The large-sample test statistics are smaller than their small-sample counterparts due to the heteroskedasticity present in the data. While the decisions of these tests are unaffected by the choice of covariance estimator, this will not always be the case.

Theorem 3.17 (Misspecified Regression). *Under assumptions 3.1 and 3.7 - 3.9 through , if \mathbf{X} can be partitioned $[\mathbf{X}_1 \quad \mathbf{X}_2]$ where \mathbf{X}_1 correspond to included variables while \mathbf{X}_2 correspond to excluded variables with non-zero coefficients, then*

$$\begin{aligned}\hat{\beta}_{1n} &\xrightarrow{p} \beta_1 + \Sigma_{\mathbf{x}_1 \mathbf{x}_1}^{-1} \Sigma_{\mathbf{x}_1 \mathbf{x}_2} \beta_2 \\ \hat{\beta}_1 &\xrightarrow{p} \beta_1 + \delta \beta_2\end{aligned}\tag{3.87}$$

where

$$\Sigma_{\mathbf{xx}} = \begin{bmatrix} \Sigma_{\mathbf{x}_1 \mathbf{x}_1} & \Sigma_{\mathbf{x}_1 \mathbf{x}_2} \\ \Sigma'_{\mathbf{x}_1 \mathbf{x}_2} & \Sigma_{\mathbf{x}_2 \mathbf{x}_2} \end{bmatrix}$$

The bias term, $\delta\beta_2$ is composed of two elements. The first, δ , is a matrix of regression coefficients where the j^{th} column is the probability limit of the least-squares estimator in the regression

$$\mathbf{X}_{2j} = \mathbf{X}_1\delta_j + \nu,$$

where \mathbf{X}_{2j} is the j^{th} column of \mathbf{X}_2 . The second component of the bias term is the original regression coefficients. As should be expected, larger coefficients on omitted variables lead to larger bias.

$\hat{\beta}_{1n} \xrightarrow{P} \beta_1$ under one of three conditions:

1. $\hat{\delta}_n \xrightarrow{P} \mathbf{0}$
2. $\beta_2 = \mathbf{0}$
3. The product $\hat{\delta}_n\beta_2 \xrightarrow{P} \mathbf{0}$.

β_2 has been assumed to be non-zero (if $\beta_2 = \mathbf{0}$ the model is correctly specified). $\delta_n \xrightarrow{P} \mathbf{0}$ only if the regression coefficients of \mathbf{X}_2 on \mathbf{X}_1 are zero, which requires that the omitted and included regressors to be uncorrelated (\mathbf{X}_2 lies in the null space of \mathbf{X}_1). This assumption should be considered implausible in most applications and $\hat{\beta}_{1n}$ is biased and inconsistent, in general. Note that certain classes of regressors that are mutually orthogonal by design and can be safely omitted.¹⁶ Finally, if both δ and β_2 are non-zero, the product could be zero, although, without a very peculiar specification and a careful selection of regressors, this possibility should be considered unlikely.

Alternatively, consider the case where some irrelevant variables are included. The correct model specification is

$$Y_i = \mathbf{X}_{1,i}\beta_1 + \varepsilon_i$$

and the model estimated is

$$Y_i = \mathbf{X}_{1,i}\beta_1 + \mathbf{X}_{2,i}\beta_2 + \varepsilon_i$$

As long as the assumptions of the asymptotic framework are satisfied, the least-squares estimator is consistent under theorem 3.12 and

$$\hat{\beta}_n \xrightarrow{P} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \mathbf{0} \end{bmatrix}$$

If the errors are homoskedastic, the variance of $\sqrt{n}(\hat{\beta}_n - \beta)$ is $\sigma^2\Sigma_{\mathbf{XX}}^{-1}$ where $\mathbf{X} = [\mathbf{X}_1 \quad \mathbf{X}_2]$. The variance of $\hat{\beta}_{1n}$ is the upper left k_1 by k_1 block of $\sigma^2\Sigma_{\mathbf{XX}}^{-1}$. Using the partitioned inverse,

$$\Sigma_{\mathbf{XX}}^{-1} = \begin{bmatrix} \Sigma_{\mathbf{X}_1\mathbf{X}_1}^{-1} + \Sigma_{\mathbf{X}_1\mathbf{X}_1}^{-1}\Sigma_{\mathbf{X}_1\mathbf{X}_2}\mathbf{M}_1\Sigma'_{\mathbf{X}_1\mathbf{X}_2}\Sigma_{\mathbf{X}_1\mathbf{X}_1}^{-1} & -\Sigma_{\mathbf{X}_1\mathbf{X}_1}^{-1}\Sigma_{\mathbf{X}_1\mathbf{X}_2}\mathbf{M}_1 \\ \mathbf{M}_1\Sigma'_{\mathbf{X}_1\mathbf{X}_2}\Sigma_{\mathbf{X}_1\mathbf{X}_1}^{-1} & \Sigma_{\mathbf{X}_2\mathbf{X}_2}^{-1} + \Sigma_{\mathbf{X}_2\mathbf{X}_2}^{-1}\Sigma'_{\mathbf{X}_1\mathbf{X}_2}\mathbf{M}_2\Sigma_{\mathbf{X}_1\mathbf{X}_2}\Sigma_{\mathbf{X}_2\mathbf{X}_2}^{-1} \end{bmatrix}$$

¹⁶Safely in terms of consistency of estimated parameters. Omitting variables will cause the estimated variance to be inconsistent.

where

$$\mathbf{M}_1 = \lim_{n \rightarrow \infty} \frac{\mathbf{X}'_2 \mathbf{M}_{\mathbf{X}_1} \mathbf{X}_2}{n}$$

$$\mathbf{M}_2 = \lim_{n \rightarrow \infty} \frac{\mathbf{X}'_1 \mathbf{M}_{\mathbf{X}_2} \mathbf{X}_1}{n}$$

and so the upper left block of the variance, $\Sigma_{\mathbf{X}_1 \mathbf{X}_1}^{-1} + \Sigma_{\mathbf{X}_1 \mathbf{X}_1}^{-1} \Sigma_{\mathbf{X}_1 \mathbf{X}_2} \mathbf{M}_1 \Sigma'_{\mathbf{X}_1 \mathbf{X}_2} \Sigma_{\mathbf{X}_1 \mathbf{X}_1}^{-1}$, must be larger than $\Sigma_{\mathbf{X}_1 \mathbf{X}_1}^{-1}$ because the second term is a quadratic form and \mathbf{M}_1 is positive semi-definite.¹⁷ Noting that $\hat{\sigma}^2$ is consistent under both the correct specification and the expanded specification, the cost of including extraneous regressors is an increase in the asymptotic variance.

In finite samples, there is a bias-variance trade-off. Fewer regressors included in a model leads to more precise estimates. Models containing more variables tend to produce coefficient estimated with less bias. Additionally, if relevant variables are omitted then $\hat{\sigma}^2$ is larger than it would be if all relevant variables are included, and so the estimated parameter variance, $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$ is also larger. Asymptotically, only the bias remains as it is of a higher order than variance (scaling $\hat{\beta}_n - \beta$ by \sqrt{n} , the bias is exploding while the variance is constant), and so when the sample size is large and estimates are precise, a larger model should be preferred to a smaller model. In cases where the sample size is small, there is a justification for omitting a variable to enhance the precision of those remaining, particularly when the effect of the omitted variable is not of interest or when the excluded variable is highly correlated with one or more included variables.

3.12.2 Errors Correlated with Regressors

Bias can arise from sources other than omitted variables. Consider the case where \mathbf{X} is measured with noise and define $\tilde{\mathbf{X}}_i = \mathbf{X}_i + \eta_i$ where $\tilde{\mathbf{X}}_i$ is a noisy proxy for \mathbf{X}_i , the “true” (unobserved) regressor, and η_i is an i.i.d. mean 0 noise process which is independent of \mathbf{X} and ε with finite second moments $\Sigma_{\eta\eta}$. The OLS estimator,

$$\hat{\beta}_n = \left(\frac{\tilde{\mathbf{X}}'\tilde{\mathbf{X}}}{n} \right)^{-1} \frac{\tilde{\mathbf{X}}'\mathbf{y}}{n} \quad (3.88)$$

$$= \left(\frac{(\mathbf{X} + \eta)'(\mathbf{X} + \eta)}{n} \right)^{-1} \frac{(\mathbf{X} + \eta)'\mathbf{y}}{n} \quad (3.89)$$

$$= \left(\frac{\mathbf{X}'\mathbf{X}}{n} + \frac{\mathbf{X}'\eta}{n} + \frac{\eta'\mathbf{X}}{n} + \frac{\eta'\eta}{n} \right)^{-1} \frac{(\mathbf{X} + \eta)'\mathbf{y}}{n} \quad (3.90)$$

$$= \left(\frac{\mathbf{X}'\mathbf{X}}{n} + \frac{\mathbf{X}'\eta}{n} + \frac{\eta'\mathbf{X}}{n} + \frac{\eta'\eta}{n} \right)^{-1} \left(\frac{\mathbf{X}'\mathbf{y}}{n} + \frac{\eta'\mathbf{y}}{n} \right) \quad (3.91)$$

will be biased downward. To understand the source of the bias, consider the behavior, under the asymptotic assumptions, of

¹⁷Both \mathbf{M}_1 and \mathbf{M}_2 are covariance matrices of the residuals of regressions of \mathbf{x}_2 on \mathbf{x}_1 and \mathbf{x}_1 on \mathbf{x}_2 respectively.

$$\begin{aligned}\frac{\mathbf{X}'\mathbf{X}}{n} &\xrightarrow{p} \Sigma_{\mathbf{XX}} \\ \frac{\mathbf{X}'\boldsymbol{\eta}}{n} &\xrightarrow{p} \mathbf{0} \\ \frac{\boldsymbol{\eta}'\boldsymbol{\eta}}{n} &\xrightarrow{p} \Sigma_{\boldsymbol{\eta}\boldsymbol{\eta}} \\ \frac{\mathbf{X}'\mathbf{y}}{n} &\xrightarrow{p} \Sigma_{\mathbf{XX}}\beta \\ \frac{\boldsymbol{\eta}'\mathbf{y}}{n} &\xrightarrow{p} \mathbf{0}\end{aligned}$$

so

$$\left(\frac{\mathbf{X}'\mathbf{X}}{n} + \frac{\mathbf{X}'\boldsymbol{\eta}}{n} + \frac{\boldsymbol{\eta}'\mathbf{X}}{n} + \frac{\boldsymbol{\eta}'\boldsymbol{\eta}}{n} \right)^{-1} \xrightarrow{p} (\Sigma_{\mathbf{XX}} + \Sigma_{\boldsymbol{\eta}\boldsymbol{\eta}})^{-1}$$

and

$$\hat{\boldsymbol{\beta}}_n \xrightarrow{p} (\Sigma_{\mathbf{XX}} + \Sigma_{\boldsymbol{\eta}\boldsymbol{\eta}})^{-1}\Sigma_{\mathbf{XX}}\beta.$$

If $\Sigma_{\boldsymbol{\eta}\boldsymbol{\eta}} \neq \mathbf{0}$, then $\hat{\boldsymbol{\beta}}_n \not\xrightarrow{p} \boldsymbol{\beta}$ and the estimator is inconsistent.

The OLS estimator is also biased in the case where $n^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \not\xrightarrow{p} \mathbf{0}_k$, which arises in situations with *endogeneity*. In these cases, \mathbf{x}_i and $\boldsymbol{\varepsilon}_i$ are simultaneously determined and correlated. This correlation results in a biased estimator since $\hat{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta} + \Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{X}\boldsymbol{\varepsilon}}$ where $\Sigma_{\mathbf{X}\boldsymbol{\varepsilon}}$ is the limit of $n^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$. The classic example of endogeneity is simultaneous equation models although many situations exist where the innovation may be correlated with one or more regressors; omitted variables can be considered a special case of endogeneity by reformulating the model.

The solution to this problem is to find an instrument, \mathbf{z}_i , which is correlated with the endogenous variable, \mathbf{x}_i , but uncorrelated with $\boldsymbol{\varepsilon}_i$. Intuitively, the endogenous portions of \mathbf{x}_i can be annihilated by regressing \mathbf{x}_i on \mathbf{z}_i and using the fit values. This procedure is known as instrumental variable (IV) regression in the case where the number of \mathbf{z}_i variables is the same as the number of \mathbf{x}_i variables and two-stage least squares (2SLS) when the size of \mathbf{z}_i is larger than k .

Define \mathbf{z}_i as a vector of exogenous variables where \mathbf{z}_i may contain any of the variables in \mathbf{x}_i which are exogenous. However, all endogenous variables – those correlated with the error – must be excluded.

First, a few assumptions must be reformulated.

Assumption 3.11 (IV Stationary Ergodicity). $\{(\mathbf{Z}_i, \mathbf{X}_i, \boldsymbol{\varepsilon}_i)\}$ is a strictly stationary and ergodic sequence.

Assumption 3.12 (IV Rank). $E[\mathbf{Z}'_i \mathbf{X}_i] = \Sigma_{\mathbf{ZX}}$ is nonsingular and finite.

Assumption 3.13 (IV Martingale Difference). $\{\mathbf{Z}'_i \boldsymbol{\varepsilon}_i, \mathcal{F}_i\}$ is a martingale difference sequence,

$$E\left[\left(Z_{j,i}\boldsymbol{\varepsilon}_i\right)^2\right] < \infty, j = 1, 2, \dots, k, i = 1, 2 \dots$$

and $\mathbf{S} = V[n^{-\frac{1}{2}}\mathbf{Z}'\boldsymbol{\varepsilon}]$ is finite and non singular.

Assumption 3.14 (IV Moment Existence). $E[X_{ji}^4] < \infty$ and $E[Z_{ji}^4] < \infty$, $j = 1, 2, \dots, k$, $i = 1, 2, \dots$ and $E[\varepsilon_i^2] = \sigma^2 < \infty$, $i = 1, 2, \dots$

These four assumptions are nearly identical to the four used to establish the asymptotic normality of the OLS estimator. The IV estimator is defined

$$\hat{\beta}_n^{IV} = \left(\frac{\mathbf{Z}'\mathbf{X}}{n} \right)^{-1} \frac{\mathbf{Z}'\mathbf{y}}{n} \quad (3.92)$$

where the n term is present to describe the number of observations used in the IV estimator. The asymptotic properties are easy to establish and are virtually identical to those of the OLS estimator.

Theorem 3.18 (Consistency of the IV Estimator). *Under assumptions 3.1 and 3.11-3.13, the IV estimator is consistent,*

$$\hat{\beta}_n^{IV} \xrightarrow{P} \beta$$

and asymptotically normal

$$\sqrt{n}(\hat{\beta}_n^{IV} - \beta) \xrightarrow{d} N(0, \Sigma_{ZX}^{-1} \ddot{\mathbf{S}} \Sigma_{ZX}^{-1}) \quad (3.93)$$

where $\Sigma_{ZX} = E[\mathbf{x}'_i \mathbf{z}_i]$ and $\ddot{\mathbf{S}} = V[n^{-1/2} \mathbf{Z}' \varepsilon]$.

Additionally, consistent estimators are available for the components of the asymptotic variance.

Theorem 3.19 (Asymptotic Normality of the IV Estimator). *Under assumptions 3.1 and 3.11 - 3.14,*

$$\hat{\Sigma}_{ZX} = n^{-1} \mathbf{Z}' \mathbf{X} \xrightarrow{P} \Sigma_{ZX} \quad (3.94)$$

$$\hat{\mathbf{S}} = n^{-1} \sum_{i=1}^n \varepsilon_i^2 \mathbf{z}'_i \mathbf{z}_i \xrightarrow{P} \ddot{\mathbf{S}} \quad (3.95)$$

and

$$\hat{\Sigma}_{ZX}^{-1} \hat{\mathbf{S}} \hat{\Sigma}_{ZX}^{-1} \xrightarrow{P} \Sigma_{ZX}^{-1} \ddot{\mathbf{S}} \Sigma_{ZX}^{-1} \quad (3.96)$$

The asymptotic variance can be easily computed from

$$\begin{aligned} \hat{\Sigma}_{ZX}^{-1} \hat{\mathbf{S}} \hat{\Sigma}_{ZX}^{-1} &= N(\mathbf{Z}' \mathbf{X})^{-1} \left(\sum_{i=1}^n \hat{\varepsilon}_i^2 \mathbf{z}'_i \mathbf{z}_i \right) (\mathbf{X}' \mathbf{Z})^{-1} \\ &= N(\mathbf{Z}' \mathbf{X})^{-1} (\mathbf{Z}' \hat{\mathbf{E}} \mathbf{Z}) (\mathbf{X}' \mathbf{Z})^{-1} \end{aligned} \quad (3.97)$$

where $\hat{\mathbf{E}} = \text{diag}(\hat{\varepsilon}_1^2, \dots, \hat{\varepsilon}_n^2)$ is a matrix with the estimated residuals squared along its diagonal.

IV estimators have one further complication beyond those of OLS. Assumption 3.8 requires the rank of $\mathbf{Z}' \mathbf{X}$ to be full (k), and so \mathbf{z}_i must be correlated with \mathbf{x}_i . Moreover, since the asymptotic variance depends on Σ_{ZX}^{-1} , even variables with non-zero correlation may produce imprecise estimates, especially if the correlation is low. Instruments must be carefully chosen, although substantially deeper treatment is beyond the scope of this course. Fortunately, IV estimators are infrequently needed in financial econometrics.

3.12.3 Monte Carlo: The effect of instrument correlation

While IV estimators are not often needed with financial data¹⁸, the problem of endogeneity is severe and it is important to be aware of the consequences and pitfalls of using IV estimators.¹⁹ To understand this problem, consider a simple Monte Carlo. The regressor (X_i), the instrument (Z_i) and the error are all drawn from a multivariate normal with the covariance matrix,

$$\begin{bmatrix} X_i \\ Z_i \\ \varepsilon_i \end{bmatrix} \sim N \left(\mathbf{0}, \begin{bmatrix} 1 & \rho_{xz} & \rho_{x\varepsilon} \\ \rho_{xz} & 1 & 0 \\ \rho_{x\varepsilon} & 0 & 1 \end{bmatrix} \right).$$

Throughout the experiment, $\rho_{x\varepsilon} = 0.4$ and ρ_{xz} is varied from 0 to .9. 200 data points were generated from

$$Y_i = \beta_1 X_i + \varepsilon_i$$

where $\beta_1 = 1$. It is straightforward to show that $E[\hat{\beta}] = 1 + \rho_{x\varepsilon}$ and that $\hat{\beta}_n^{\text{IV}} \xrightarrow{P} 1$ as long as $\rho_{xz} \neq 0$. 10,000 replications were generated and the IV estimators were computed

$$\hat{\beta}_n^{\text{IV}} = (\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{y}).$$

Figure 3.5 contains kernel density plots of the instrumental variable estimator for ρ_{xz} of .2, .4, .6 and .8. When the correlation between the instrument and X is low, the distribution is dispersed (exhibiting a large variance). As the correlation increases, the variance decreases and the distribution become increasingly normal. This experiment highlights two fundamental problems with IV estimators: they have large variance when no “good instruments” – highly correlated with \mathbf{x}_i by uncorrelated with ε_i – are available and the finite-sample distribution of IV estimators may be poorly approximated a normal.

3.12.4 Heteroskedasticity

Assumption 3.7 does not require data to be homoskedastic, which is useful since heteroskedasticity is the rule rather than the exception in financial data. If the data are homoskedastic, the asymptotic covariance of $\hat{\beta}$ can be consistently estimated by

$$\hat{\mathbf{S}} = \hat{\sigma}^2 \left(\frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1}$$

Heteroskedastic errors require the use of a more complicated covariance estimator, and the asymptotic variance can be consistently estimated using

¹⁸IV estimators are most common in corporate finance when examining executive compensation and company performance.

¹⁹The intuition behind IV estimators is generally applicable to 2SLS.

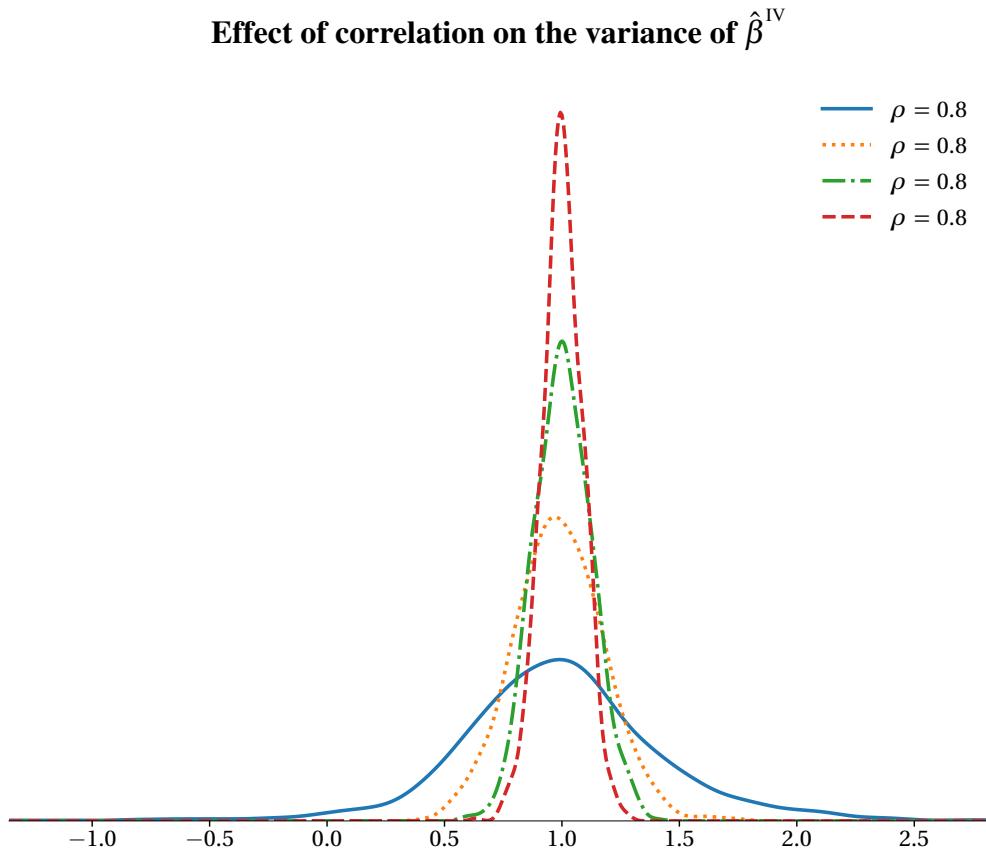


Figure 3.5: Kernel density of the instrumental variable estimator $\hat{\beta}_n^{\text{IV}}$ with varying degrees of correlation between the endogenous variable and the instrument. Increasing the correlation between the instrument and the endogenous variable leads to a large decrease in the variance of the estimated parameter ($\beta = 1$). When the correlation is small (.2), the distribution has a large variance and is not well approximated by a normal random variable.

$$\begin{aligned}
 \hat{\Sigma}_{\mathbf{X}\mathbf{X}}^{-1} \hat{\mathbf{S}} \hat{\Sigma}_{\mathbf{X}\mathbf{X}}^{-1} &= \left(\frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1} \left(\frac{\sum_{i=1}^n \hat{\varepsilon}_i^2 \mathbf{x}_i' \mathbf{x}_i}{n} \right) \left(\frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1} \\
 &= n (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{i=1}^n \hat{\varepsilon}_i^2 \mathbf{x}_i' \mathbf{x}_i \right) (\mathbf{X}'\mathbf{X})^{-1} \\
 &= n (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}'\hat{\mathbf{E}}\mathbf{X}) (\mathbf{X}'\mathbf{X})^{-1}
 \end{aligned} \tag{3.98}$$

where $\hat{\mathbf{E}} = \text{diag}(\hat{\varepsilon}_1^2, \dots, \hat{\varepsilon}_n^2)$ is a matrix with the estimated residuals squared along its diagonal.

Faced with two covariance estimators, one which is consistent under minimal assumptions and one which requires an additional, often implausible assumption, it may be tempting to rely exclusively on the robust estimator. This covariance estimator is known as the White heteroskedasticity-consistent covariance estimator and standard errors computed using eq. (3.98) are called heteroskedasticity-robust standard errors or White standard errors (White, 1980). Using a heteroskedasticity-consistent

estimator when not needed (homoskedastic data) results in test statistics that have worse small-sample properties. In small samples, hypothesis tests are more likely to have size distortions and so using 5% critical values may lead to rejection of the null 10% or more of the time when the null is true. On the other hand, using an inconsistent estimator of the parameter covariance – assuming homoskedasticity when the data are not – produces tests with size distortions, even asymptotically.

White (1980) also provides a test to determine if a heteroskedasticity robust covariance estimator is required. Each term in the heteroskedasticity-consistent estimator takes the form

$$\boldsymbol{\varepsilon}_i^2 \mathbf{x}'_i \mathbf{x}_i = \begin{bmatrix} \boldsymbol{\varepsilon}_i^2 x_{1,i}^2 & \boldsymbol{\varepsilon}_i^2 x_{1,i} x_{2,i} & \dots & \boldsymbol{\varepsilon}_i^2 x_{1,i} x_{kn} \\ \boldsymbol{\varepsilon}_i^2 x_{1,i} x_{2,i} & \boldsymbol{\varepsilon}_i^2 x_{2,i}^2 & \dots & \boldsymbol{\varepsilon}_i^2 x_{2,i} x_{kn} \\ \vdots & \vdots & \dots & \vdots \\ \boldsymbol{\varepsilon}_i^2 x_{1,i} x_{kn} & \boldsymbol{\varepsilon}_i^2 x_{2,i} x_{kn} & \dots & \boldsymbol{\varepsilon}_i^2 x_{kn}^2 \end{bmatrix},$$

and so, if $E[\boldsymbol{\varepsilon}_i^2 x_{jn} x_{ln}] = E[\boldsymbol{\varepsilon}_i^2] E[x_{jn} x_{ln}]$, for all j and l , then the heteroskedasticity robust and the standard estimator will both consistently estimate the asymptotic variance of $\hat{\beta}$. White's test is formulated as a regression of *squared* estimated residuals on all unique squares and cross products of \mathbf{x}_i . Suppose the original regression specification is

$$Y_i = \beta_1 + \beta_2 X_{1,i} + \beta_3 X_{2,i} + \boldsymbol{\varepsilon}_i.$$

White's test uses an auxiliary regression of $\hat{\boldsymbol{\varepsilon}}_i^2$ on the squares and cross-products of all regressors, $\{1, X_{1,i}, X_{2,i}, X_{1,i}^2, X_{2,i}^2, X_{1,i}X_{2,i}\}$:

$$\hat{\boldsymbol{\varepsilon}}_i^2 = \delta_1 + \delta_2 X_{1,i} + \delta_3 X_{2,i} + \delta_4 X_{1,i}^2 + \delta_5 X_{2,i}^2 + \delta_6 X_{1,i} X_{2,i} + \eta_i. \quad (3.99)$$

The null hypothesis tested is $H_0 : \delta_j = 0$, $j > 1$, and the test statistic can be computed using nR^2 where the centered R^2 is from the model in eq. (3.99). Recall that nR^2 is an LM test of the null that all coefficients except the intercept are zero and has an asymptotic χ_v^2 where v is the number of restrictions – the same as the number of regressors excluding the constant. If the null is rejected, a heteroskedasticity robust covariance estimator is required.

Algorithm 3.8 (White's Test).

1. Fit the model $Y_i = \mathbf{X}_i \beta + \boldsymbol{\varepsilon}_i$
2. Construct the fit residuals $\hat{\boldsymbol{\varepsilon}}_i = Y_i - \mathbf{X}_i \hat{\beta}$
3. Construct the auxiliary regressors \mathbf{Z}_i where the $k(k+1)/2$ elements of \mathbf{z}_i are computed from $X_{i,o} X_{i,p}$ for $o = 1, 2, \dots, k$, $p = o, o+1, \dots, k$.
4. Estimate the auxiliary regression $\hat{\boldsymbol{\varepsilon}}_i^2 = \mathbf{Z}_i \gamma + \eta_i$
5. Compute White's Test statistic as nR^2 where the R^2 is from the auxiliary regression and compare to the critical value at size α from a $\chi_{k(k+1)/2-1}^2$.

3.12.5 Example: White's test on the FF data

White's heteroskedasticity test is implemented using the estimated residuals, $\hat{\varepsilon}_i = Y_i - \mathbf{x}'_i \hat{\beta}$, by regressing the estimated residuals squared on all unique cross products of the regressors. The primary model fit is

$$BH_i^e = \beta_1 + \beta_2 VWM_i^e + \beta_3 SMB_i + \beta_4 HML_i + \beta_5 MOM_i + \varepsilon_i.$$

and the auxiliary model is specified

$$\begin{aligned} \hat{\varepsilon}_i^2 = & \delta_1 + \delta_2 VWM_i^e + \delta_3 SMB_i + \delta_4 HML_i + \delta_5 MOM_i + \delta_6 (VWM_i^e)^2 + \delta_7 VWM_i^e SMB_i \\ & + \delta_8 VWM_i^e HML_i + \delta_9 VWM_i^e MOM_i + \delta_{10} SMB_i^2 + \delta_{11} SMB_i HML_i \\ & + \delta_{12} SMB_i MOM_i + \delta_{13} HML_i^2 + \delta_{14} HML_i MOM_i + \delta_{15} MOM_i^2 + \eta_i \end{aligned}$$

Estimating this regression produces an R^2 of 10.9% and $nR^2 = 74.8$, which has an asymptotic χ^2_{14} distribution (14 regressors, excluding the constant). The p-value of this test statistic is 0.000, and the null of homoskedasticity is strongly rejected.

3.12.6 Generalized Least Squares

An alternative to modeling heteroskedastic data is to transform the data so that it is homoskedastic using generalized least squares (GLS). GLS extends OLS to allow for arbitrary weighting matrices. The GLS estimator of β is defined

$$\hat{\beta}^{GLS} = (\mathbf{X}' \mathbf{W}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W}^{-1} \mathbf{y}, \quad (3.100)$$

for some positive definite matrix \mathbf{W} . Without any further assumptions or restrictions on \mathbf{W} , $\hat{\beta}^{GLS}$ is unbiased under the same conditions as $\hat{\beta}$, and the variance of $\hat{\beta}$ can be easily shown to be

$$(\mathbf{X}' \mathbf{W}^{-1} \mathbf{X})^{-1} (\mathbf{X}' \mathbf{W}^{-1} \mathbf{V} \mathbf{W}^{-1} \mathbf{X}) (\mathbf{X}' \mathbf{W}^{-1} \mathbf{X})^{-1}$$

where \mathbf{V} is the n by n covariance matrix of ε .

The full value of GLS is only realized when \mathbf{W} is wisely chosen. Suppose that the data are heteroskedastic but not serial correlated,²⁰ and so

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \quad (3.101)$$

where $V[\varepsilon_i | \mathbf{X}] = \sigma_i^2$ and therefore heteroskedastic. Further, assume σ_i^2 is known. Returning to the small-sample assumptions, choosing $\mathbf{W} \propto V(\varepsilon | \mathbf{X})^{21}$, the GLS estimator will be efficient.

Assumption 3.15 (Error Covariance). $\mathbf{V} = V[\varepsilon | \mathbf{X}]$

Setting $\mathbf{W} = \mathbf{V}$, the GLS estimator is BLUE.

²⁰Serial correlation is ruled out by assumption 3.9.

²¹ \propto is the mathematical symbol for “proportional to”.

Theorem 3.20 (Variance of $\hat{\beta}^{\text{GLS}}$). *Under assumptions 3.1 - 3.3 and 3.15,*

$$\mathbf{V}[\hat{\beta}^{\text{GLS}} | \mathbf{X}] = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1}$$

and $\mathbf{V}[\hat{\beta}^{\text{GLS}} | \mathbf{X}] \leq \mathbf{V}[\tilde{\beta} | \mathbf{X}]$ where $\tilde{\beta} = \mathbf{C}\mathbf{y}$ is any other linear unbiased estimator with $E[\tilde{\beta}] = \beta$

To understand the intuition behind this result, note that the GLS estimator can be expressed as an OLS estimator using transformed data. Returning to the model in eq. (3.101), and pre-multiplying by $\mathbf{W}^{-\frac{1}{2}}$,

$$\begin{aligned} \mathbf{W}^{-\frac{1}{2}}\mathbf{y} &= \mathbf{W}^{-\frac{1}{2}}\mathbf{X}\beta + \mathbf{W}^{-\frac{1}{2}}\varepsilon \\ \tilde{\mathbf{y}} &= \tilde{\mathbf{X}}\beta + \tilde{\varepsilon} \end{aligned}$$

and so

$$\begin{aligned} \hat{\beta} &= (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})\tilde{\mathbf{X}}'\tilde{\mathbf{y}} \\ &= \left(\mathbf{X}'\mathbf{W}^{-\frac{1}{2}}\mathbf{W}^{-\frac{1}{2}}\mathbf{X}\right)\mathbf{X}'\mathbf{W}^{-\frac{1}{2}}\mathbf{W}^{-\frac{1}{2}}\mathbf{y} \\ &= (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})\mathbf{X}'\mathbf{W}^{-1}\mathbf{y} \\ &= \hat{\beta}^{\text{GLS}}. \end{aligned}$$

In the original model, $\mathbf{W} = \mathbf{V}[\varepsilon | \mathbf{X}]$, and so $\mathbf{V}[\mathbf{W}^{-\frac{1}{2}}\varepsilon | \mathbf{X}] = \mathbf{W}^{-\frac{1}{2}}\mathbf{W}\mathbf{W}^{-\frac{1}{2}} = \mathbf{I}_n$. $\tilde{\varepsilon}$ is homoskedastic and uncorrelated and the transformed model satisfies the assumption of the Gauss-Markov theorem (theorem 3.3).

This result is only directly applicable under the small-sample assumptions and then only if $\mathbf{V}[\varepsilon | \mathbf{X}]$ is known *a priori*. In practice, neither is true: data are not congruent with the small-sample assumptions and $\mathbf{V}[\varepsilon | \mathbf{X}]$ is never known. The feasible GLS (FGLS) estimator solves these two issues, although the efficiency gains of FGLS have only asymptotic justification. Suppose that $\mathbf{V}[\varepsilon | \mathbf{X}] = \omega_1 + \omega_2 x_{1,i} + \dots + \omega_{k+1} x_{kn}$ where ω_j are unknown. The FGLS procedure provides a method to estimate these parameters and implement a feasible GLS estimator.

The FGLS procedure is described in the following algorithm.

Algorithm 3.9 (Feasible GLS Estimation).

1. Estimate $\hat{\beta}$ using OLS.
2. Using the estimated residuals, $\hat{\varepsilon} = \mathbf{y} - \mathbf{X}\hat{\beta}$, estimate an auxiliary model by regressing the squared residual on the variables of the variance model.
3. Using the estimated variance model parameters $\hat{\omega}$, produce a fit variance matrix, $\hat{\mathbf{V}}$.
4. Compute $\tilde{\mathbf{y}} = \hat{\mathbf{V}}^{-\frac{1}{2}}\mathbf{y}$ and $\tilde{\mathbf{X}} = \hat{\mathbf{V}}^{-\frac{1}{2}}\mathbf{X}$ compute $\hat{\beta}^{\text{FGLS}}$ using the OLS estimator on the transformed regressors and regressand.

Hypothesis testing can be performed on $\hat{\beta}^{\text{FGLS}}$ using the standard test statistics with the FGLS variance estimator,

$$\tilde{\sigma}^2(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1} = \tilde{\sigma}^2(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}$$

where $\tilde{\sigma}^2$ is the sample variance of the FGLS regression errors ($\tilde{\epsilon} = \tilde{\mathbf{y}} - \tilde{\mathbf{X}}\hat{\beta}^{\text{FGLS}}$).

While FGLS is only formally asymptotically justified, FGLS estimates are often much more precise in finite samples, especially if the data is very heteroskedastic. Estimator accuracy improves the most when some observations have a vastly larger variance than others. The OLS estimator gives these observations too much weight, inefficiently exploiting the information in the remaining observations. FGLS, even when estimated with a diagonal weighting matrix that may be slightly misspecified, can produce substantially more precise estimates.²²

3.12.6.1 Monte Carlo: A simple GLS

A simple Monte Carlo was designed to demonstrate the gains of GLS. The observed data are generated according to

$$Y_i = X_i + X_i^\alpha \epsilon_i$$

where X_i is i.i.d. $U(0,1)$ and ϵ_i is standard normal. α takes the values of 0.8, 1.6, 2.8 and 4. When α is low the data are approximately homoskedastic. As α increases the data are increasingly heteroskedastic and the probability of producing a few residuals with small variances increases. The OLS and (infeasible) GLS estimators were fit to the data and figure 3.6 contains kernel density plots of $\hat{\beta}$ and $\hat{\beta}^{GLS}$.

When α is small, the OLS and GLS parameter estimates have similar variances, indicated by the similarity in distribution. As α increases, the GLS estimator becomes very precise which is due to GLS's reweighting of the data by the inverse of its variance. In effect, observations with the smallest errors become very influential in determining $\hat{\beta}$. This is the general principle behind GLS: let the data points which are most precise about the unknown parameters have the most influence.

3.12.7 Example: GLS in the Factor model

Even if it is unreasonable to assume that the entire covariance structure of the residuals can be correctly specified in the auxiliary regression, GLS estimates are often much more precise than OLS estimates. Consider the regression of BH^e on the four factors and a constant. The OLS estimates are identical to those previously presented and the GLS estimates will be computed using the estimated variances from White's test. Define

$$\hat{\mathbf{V}} = \text{diag}(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \dots, \hat{\sigma}_n^2)$$

where $\hat{\sigma}_i^2$ is the fit value from the auxiliary regression in White's test that included only the squares of the explanatory variables. Coefficients were estimated by regressing $\tilde{\mathbf{y}}$ on $\tilde{\mathbf{X}}$ where

$$\tilde{\mathbf{y}} = \hat{\mathbf{V}}^{-\frac{1}{2}}\mathbf{y}$$

²²If the model for the conditional variance of ϵ_i is misspecified in an application of FGLS, the resulting estimator is not asymptotically efficient and a heteroskedasticity robust covariance estimator is required.

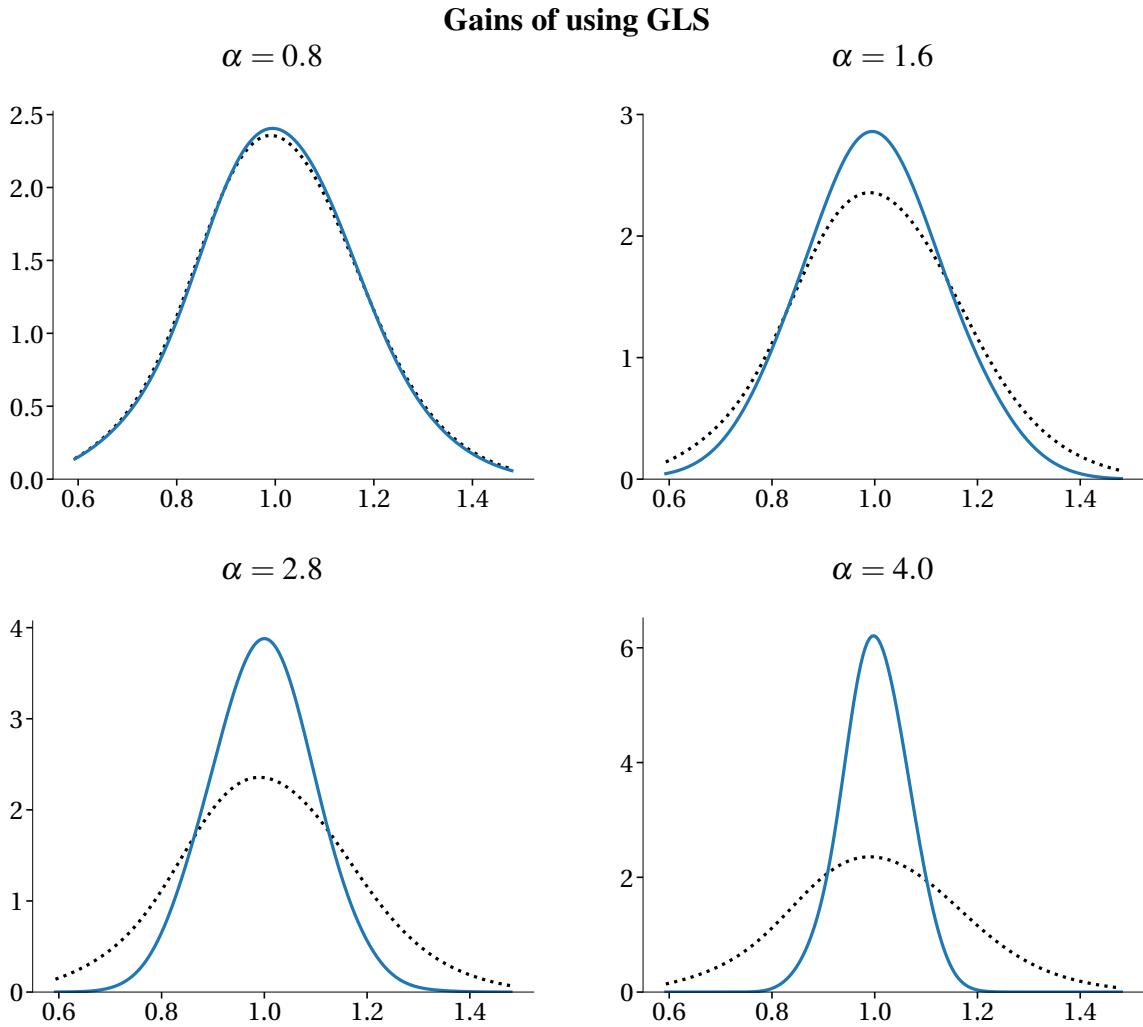


Figure 3.6: The four plots show the gains to using the GLS estimator on heteroskedastic data. The data were generated according to $Y_i = X_i + X_i^\alpha \varepsilon_i$ where X_i is i.i.d. uniform and ε_i is standard normal. For large α , the GLS estimator is substantially more efficient than the OLS estimator. However, the intuition behind the result is not that high variance residuals have been down-weighted, but that low variance residuals, some with very low variances, have been up-weighted to produce an accurate fit.

$$\tilde{\mathbf{X}} = \hat{\mathbf{V}}^{-\frac{1}{2}} \mathbf{X}$$

and $\hat{\beta}^{GLS} = (\tilde{\mathbf{X}}' \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}' \tilde{\mathbf{y}}$. $\hat{\varepsilon}^{GLS} = \mathbf{y} - \mathbf{X} \hat{\beta}^{GLS}$ are computed from the original data using the GLS estimate of β , and the variance of the GLS estimator can be computed using

$$(\tilde{\mathbf{X}}' \tilde{\mathbf{X}})^{-1} (\tilde{\mathbf{X}}' \hat{\mathbf{E}} \tilde{\mathbf{X}})^{-1} (\tilde{\mathbf{X}}' \tilde{\mathbf{X}})^{-1}.$$

where $\hat{\mathbf{E}}$ is a diagonal matrix with the estimated residuals squared, $(\hat{\varepsilon}_i^{GLS})^2$, from the GLS procedure along its diagonal. Table 3.9 contains the estimated parameters, t-stats and p-values using both the

	OLS				GLS			
	$\hat{\beta}$	s.e.($\hat{\beta}$)	t-stat	p-values	$\hat{\beta}^{GLS}$	s.e.($\hat{\beta}^{GLS}$)	t-stats	p-values
Constant	-0.09	0.04	-1.99	0.05	-0.09	0.04	-2.26	0.02
VWM^e	1.08	0.01	93.5	0.00	1.08	0.01	101.6	0.00
SMB	0.00	0.02	0.11	0.91	-0.00	0.02	-0.19	0.85
HML	0.76	0.02	36.4	0.00	0.73	0.02	39.3	0.00
MOM	-0.04	0.01	-2.63	0.01	-0.04	0.01	-3.06	0.00

Table 3.9: OLS and GLS parameter estimates and t -stats. t -stats indicate that the GLS parameter estimates are more precise.

OLS and the GLS estimates. The GLS estimation procedure appears to provide more precise estimates and inference. The difference in precision is particularly large for SMB .

3.13 Model Selection and Specification Checking

Econometric problems often begin with a variable whose dynamics are of interest and a relatively large set of candidate explanatory variables. The process by which the set of regressors is reduced is known as model selection or building.

Model building inevitably reduces to balancing two competing considerations: congruence and parsimony. A congruent model is one that captures all of the variation in the data explained by the regressors. Obviously, including all of the regressors and all functions of the regressors should produce a congruent model. However, this is also an infeasible procedure since there are infinitely many functions of even a single regressor. Parsimony dictates that the model should be as simple as possible and so models with fewer regressors are favored. The ideal model is the *parsimonious congruent* model that contains all variables necessary to explain the variation in the regressand and nothing else.

Model selection is as much a black art as science and some lessons can only be taught through experience. One principle that should be universally applied when selecting a model is to rely on economic theory and, failing that, common sense. The simplest method to select a poorly performing model is to try any and all variables, a process known as data snooping that is capable of producing a model with an arbitrarily high R^2 even if there is no relationship between the regressand and the regressors.

There are a few variable selection methods which can be examined for their properties. These include:

- General to Specific modeling (GtS)
- Specific to General modeling (StG)
- Information criteria (IC)
- Cross-validation

3.13.1 Model Building

3.13.1.1 General to Specific

General to specific (GtS) model building begins by estimating the largest model that can be justified by economic theory (and common sense). This model is then pared down to produce the smallest model that remains congruent with the data. The simplest version of GtS begins with the complete model. If any coefficients have individual p-values less than some significance level α (usually 5 or 10%), the least significant regressor is dropped from the regression. The procedure is repeated using the remaining included regressors until all coefficients are statistically significant. In each step, the least significant regressor is removed from the model.

One drawback to this simple procedure is that variables that are correlated but relevant are often dropped. This is due to a problem known as multicollinearity and individual t -stats will be small but joint significance tests that all coefficients are simultaneously zero will strongly reject. This suggests using joint hypothesis tests to pare the general model down to the specific one. While theoretically attractive, the scope of possible joint hypothesis tests is vast even in a small model, and so using joint test is impractical.

GtS suffers from two additional issues. First, it will include an irrelevant variable with positive probability (asymptotically) but will never exclude a relevant variable. Second, test statistics do not have standard distributions when they are used sequentially (as is the case with any sequential model building procedure). The only viable solution to the second problem is to fit a single model, make variable inclusions and exclusion choices, and live with the result. This practice is not typically followed and most econometricians use an iterative procedure despite the problems of sequential testing.

3.13.1.2 Specific to General

Specific to General (StG) model building begins by estimating the smallest model, usually including only a constant. Variables are then added sequentially based on maximum t -stat until there is no excluded variable with a significant t -stat at some predetermined α (again, usually 5 or 10%). StG suffers from the same issues as GtS. First it will asymptotically include all relevant variables and some irrelevant ones and second, tests implemented sequentially do not have correct size. Choosing between StG and GtS is mainly user preference, although they rarely select the same model. One argument in favor of using a GtS approach is that the variance is consistently estimated in the first step of the general specification while the variance estimated in the first step of the an StG selection is too large. This leads StG processes to have t -stats that are smaller than GtS t -stats and so StG generally selects a smaller model than GtS.

3.13.1.3 Information Criteria

The third method of model selection uses Information Criteria (IC). Information Criteria reward the model for producing smaller SSE while punishing it for the inclusion of additional regressors. The two most frequently used are the Akaike Information Criterion (AIC) and Schwarz Information Criterion (SIC) or Bayesian Information Criterion (BIC).²³ Most Information Criteria are of the form

²³The BIC and SIC are the same. BIC is probably the most common name but SIC or S/BIC are also frequently encountered.

$$-2l + P$$

where l is the log-likelihood value at the parameter estimates and P is a penalty term. In the case of least squares, where the log-likelihood is not known (or needed), IC's take the form

$$\ln \hat{\sigma}^2 + P$$

where the penalty term is divided by n .

Definition 3.15 (Akaike Information Criterion (AIC)). For likelihood-based models the AIC is defined

$$AIC = -2l + 2k \quad (3.102)$$

and in its least squares application,

$$AIC = \ln \hat{\sigma}^2 + \frac{2k}{n} \quad (3.103)$$

Definition 3.16 (Schwarz/Bayesian Information Criterion (S/BIC)). For likelihood-based models the BIC (SIC) is defined

$$BIC = -2l + k \ln n \quad (3.104)$$

and in its least squares applications

$$BIC = \ln \hat{\sigma}^2 + k \frac{\ln n}{n} \quad (3.105)$$

The obvious difference between these two IC is that the AIC has a constant penalty term while the BIC has a penalty term that increases with the number of observations. The effect of the sharper penalty in the S/BIC is that for larger data sizes, the marginal increase in the likelihood (or decrease in the variance) must be greater. This distinction is subtle but important: using the BIC to select from a finite set of regressors leads to the correct model being chosen while the AIC asymptotically selects a model that includes irrelevant regressors.

Using an IC to select a model is similar to either a GtS or StG search. For example, to use an StG selection method, begin with the smallest model (usually a constant) and compute the IC for this model. Next, consider all possible univariate regressions. If any reduce the IC, extend the specification to include the variable that produced the smallest IC. Now, beginning from the selected univariate regression, estimate all bivariate regressions. Again, if any decrease the IC, choose the one which produces the smallest value. Repeat this procedure until the marginal contribution to the IC of adding any additional variable is positive (i.e., when comparing an L and $L + 1$ variable regression, including an additional variables increase the IC).

As an alternative, if the number of regressors is sufficiently small (less than 20) it is possible to try every possible combination and choose the smallest IC. This requires 2^L regressions where L is the number of available regressors (2^{20} is about 1,000,000).

3.13.1.4 Cross-validation

Cross-validation uses pseudo-out-of-sample prediction performance to assess model specification. It is most commonly used to select a preferred model from a set of candidate models, for example, the collection of models visited as part of a GtS or StG model selection process. Variables with robust predictive power should be useful both in- and out-of-sample. Cross-validation estimates parameters

using a random subset of the data and then computes the pseudo-out-of-sample SSE on the observations that were not used in estimation. This criterion rewards models include variables with good predictive power and exclude models that incorporate variables with small coefficients that do not improve out-of-sample prediction.

The mutually exclusive and exhaustive subsets used for estimation and evaluation are randomly chosen. This randomization selection is then repeatedly applied to assess the out-of-sample fit of all data points. The most common form of cross-validation used in cross-sectional analysis is as k -fold cross-validation. This method splits the data into k -equal-sized blocks where block assignment is random. Model parameters are then estimated using the data in $k - 1$ blocks, and the predictive power is evaluated on the excluded block. This leave-one-block-out strategy is then repeated for each of the remaining $k - 1$ blocks. The overall cross-validated SSE is computed from the SSE values calculated on each block held out of the estimation.

Algorithm 3.10 (k -fold Cross-validation).

1. *Split the data randomly into k -equal-sized bins*
2. *For each model $m = 1, \dots, M$ under consideration*
 - (a) *For $i = 1, \dots, k$*
 - i. *Estimate model parameters excluding the the observations in block i ,*
$$\hat{\beta}_{m,i} = \arg \min \beta_{m,i} \sum_{j=1, j \notin \mathcal{B}_i}^n (Y_j - \mathbf{x}_{m,j} \beta_{m,i})^2$$

where $\mathbf{x}_{m,\cdot}$ are the regressors included in model m and \mathcal{B}_i is the set of observation indices in block i .

 - ii. *Compute the block i SSE as $SSE_{m,i} = \sum_{j \in \mathcal{B}_i} (Y_j - \mathbf{x}_{m,j} \hat{\beta}_{m,i})^2$.*
 - (b) *Compute the overall cross-validated SSE as $SSE_{m,CV} = \sum_{i=1}^k SSE_{m,i}$.*
3. *Select the model that produces the smallest cross-validates SSE.*

3.13.2 Specification Checking

Once a model has been selected, the final step is to examine the specification, where a number of issues may arise. For example, a model may have neglected some nonlinear features in the data, a few outliers may be determining the parameter estimates, or the data may be heteroskedastic. Residuals for the basis of most specification checks, although the first step in assessing model fit is always to plot the residuals. A simple residual plot often reveals problems with a model, such as large (and generally influential) residuals or correlation among the residuals in time-series applications.

Residual Plots and Nonlinearity Plot, plot, plot. Plots of both data and residuals, while not perfect, are effective methods to detect specification problems. Most data analysis should include a plot of the initial unfiltered data where large observation or missing data are easily detected. Once

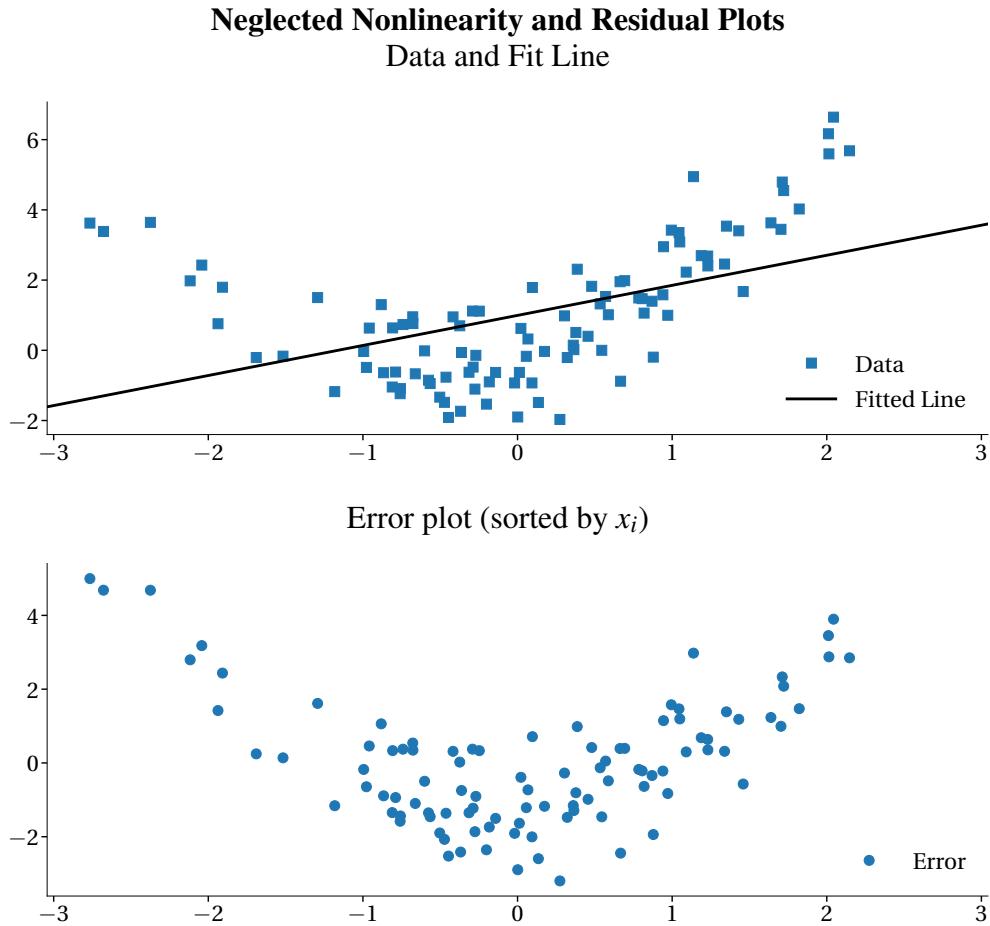


Figure 3.7: The top panel contains data generated according to $Y_i = X_i + X_i^2 + \varepsilon_i$ and a fit from a model $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$. The nonlinearity should be obvious, but is even clearer in the ordered (by X_i) residual plot where a distinct "U" shape can be seen (bottom panel).

a model has been estimated the residuals should be plotted, usually by sorting them against the ordered regressors when using cross-sectional data or against time (the observation index) in time-series applications.

To see the benefits of plotting residuals, suppose the data were generated by $Y_i = X_i + X_i^2 + \varepsilon_i$ where X_i and ε_i are i.i.d. standard normal, but an affine specification, $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$ was fit. Figure 3.7 contains plots of the data and fit lines (top panel) and errors (bottom panel). It is obvious from the data and fit line that the model is misspecified and the residual plot makes this clear. Residuals should have no discernible pattern in their mean when plotted against any variable (or function of variables) in the data set.

One statistical test for detecting neglected nonlinearity is Ramsey's RESET test. Suppose the model

$$Y_i = \mathbf{X}_i \boldsymbol{\beta} + \varepsilon_i$$

is fit and one desires to test whether there is a neglected nonlinearity present. The RESET test uses

powers of the fit data, \hat{Y}_i as additional regressors to test whether there is evidence of nonlinearity in the data.

Definition 3.17 (Ramsey's RESET Test). The RESET test is a test of the null the null $H_0 : \gamma_1 = \dots = \gamma_R = 0$ in an auxiliary regression,

$$Y_i = \mathbf{X}_i\beta + \gamma_1\hat{Y}_i^2 + \gamma_2\hat{Y}_i^3 + \dots + \gamma_R\hat{Y}_i^{R-1}\varepsilon_i$$

where \hat{Y}_i are the fit values of Y_i generated in the initial regression. The test statistic has an asymptotic χ^2_R distribution.

R is typically 1 or 2 since higher powers may produce numerical problems, imprecise estimates, and size distortions. The biggest difficulty of using a RESET test is that rejection of the null is not informative about the changes needed to the original specification.

3.13.2.1 Parameter Stability

Parameter instability is a common problem in actual data. For example, recent evidence suggests that the market β in a CAPM may be differ across up and down markets Ang, Chen, and Xing (2006). A model fit assuming the strict CAPM would be misspecified since the parameters are not constant.

There is a simple procedure to test for parameter stability if the point where the parameters changes is known. The test is specified by including a dummy for any parameter that may change and testing the coefficient on the dummy variables for constancy.

Returning to the CAPM example, the standard specification is

$$R_i^e = \beta_1 + \beta_2(R_i^M - R_i^f) + \varepsilon_i$$

where R_i^M is the return on the market, R_i^f is the return on the risk free asset and R_i^e is the excess return on the dependent asset. To test whether the slope is different when $(R_i^M - R_i^f) < 0$, define a dummy $I_i = I_{[(R_i^M - R_i^f) < 0]}$ and perform a standard test of the null $H_0 : \beta_3 = 0$ in the regression

$$R_i^e = \beta_1 + \beta_2(R_i^M - R_i^f) + \beta_3 I_i (R_i^M - R_i^f) + \varepsilon_i.$$

If the breakpoint is not known *a priori*, it is necessary to test whether there is a break in the parameter at any point in the sample. This test can be implemented by testing at every point and then examining the largest test statistic. While this is a valid procedure, the distribution of the largest test statistic is no longer χ^2 and so inference based on standard tests (and their corresponding distributions) will be misleading. This type of problem is known as a nuisance parameter problem. If the null hypothesis (that there is no break) is correct, then the value of regression coefficients after the break is not well defined. In the example above, if there is no break, then β_3 is not identified (and is a nuisance). Treatment of the issues surrounding nuisance parameters is beyond the scope of this course, but interested readers should start see Andrews and Ploberger (1994).

3.13.2.2 Rolling and Recursive Parameter Estimates

Rolling and recursive parameter estimates are useful tools for detecting parameter instability in cross-section regression of time-series data (e.g., asset returns). Rolling regression estimates use a fixed-length sample of data to estimate β and then “roll” the sampling window to produce a sequence of estimates.

Definition 3.18 (*m*-sample Rolling Regression Estimates). The *m*-sample rolling regression estimates are defined as the sequence

$$\hat{\beta}_j = \left(\sum_{i=j}^{j+m-1} \mathbf{x}'_i \mathbf{x}_i \right)^{-1} \mathbf{x}'_i Y_i \quad (3.106)$$

for $j = 1, 2, \dots, n - m + 1$.

The rolling window length should be large enough so that parameter estimates in each window are reasonably well approximated by a CLT but not so long as to smooth out any variation in β . 60-months is a common window length in applications using monthly asset price data and window lengths ranging between 3-months and 2-year are common when using daily data. The rolling regression coefficients can be visually inspected for evidence of instability, and approximate confidence intervals (based on an assumption of parameter stability) can be constructed by estimating the parameter covariance on the full sample of n observations and then scaling by n/m so that the estimated covariance is appropriate for a sample of m observations. The parameter covariance can alternatively be estimated by averaging the $n - m + 1$ covariance estimates corresponding to each sample, $\hat{\Sigma}_{\mathbf{XX},j}^{-1} \hat{\mathbf{S}}_j \hat{\Sigma}_{\mathbf{XX},j}^{-1}$, where

$$\hat{\Sigma}_{\mathbf{XX},j} = m^{-1} \sum_{i=j}^{j+m-1} \mathbf{x}'_i \mathbf{x}_i \quad (3.107)$$

and

$$\hat{\mathbf{S}}_j = m^{-1} \sum_{i=j}^{j+m-1} \hat{\epsilon}_{i,j} \mathbf{x}'_i \mathbf{x}_i \quad (3.108)$$

where $\hat{\epsilon}_{i,j} = Y_i - \mathbf{x}'_i \hat{\beta}_j$, and if the parameters are stable these methods for estimating the parameter covariance should produce similar confidence intervals.

60-month rolling regressions of the *BH* portfolio in the 4-factor model are presented in figure 3.8 where approximate confidence intervals were computed using the re-scaled full-sample parameter covariance estimate. While these confidence intervals cannot directly be used to test for parameter instability, the estimate of the loadings on the market, *SMB* and *HML* vary more than their intervals indicate these parameters should be stable.

An alternative to rolling regressions is to recursively estimate parameters which uses an expanding window of observations to estimate $\hat{\beta}$.

Definition 3.19 (Recursive Regression Estimates). Recursive regression estimates are defined as the sequence

$$\hat{\beta}_j = \left(\sum_{i=1}^j \mathbf{x}'_i \mathbf{x}_i \right)^{-1} \mathbf{x}'_i Y_i \quad (3.109)$$

for $j = l, 2, \dots, n$ where $l > k$ is the smallest window used.

Approximate confidence intervals can be computed either by re-scaling the full-sample parameter covariance or by directly estimating the parameter covariance in each recursive sample. Documenting evidence of parameter instability using recursive estimates is often more difficult than with rolling, as demonstrated in figure 3.9

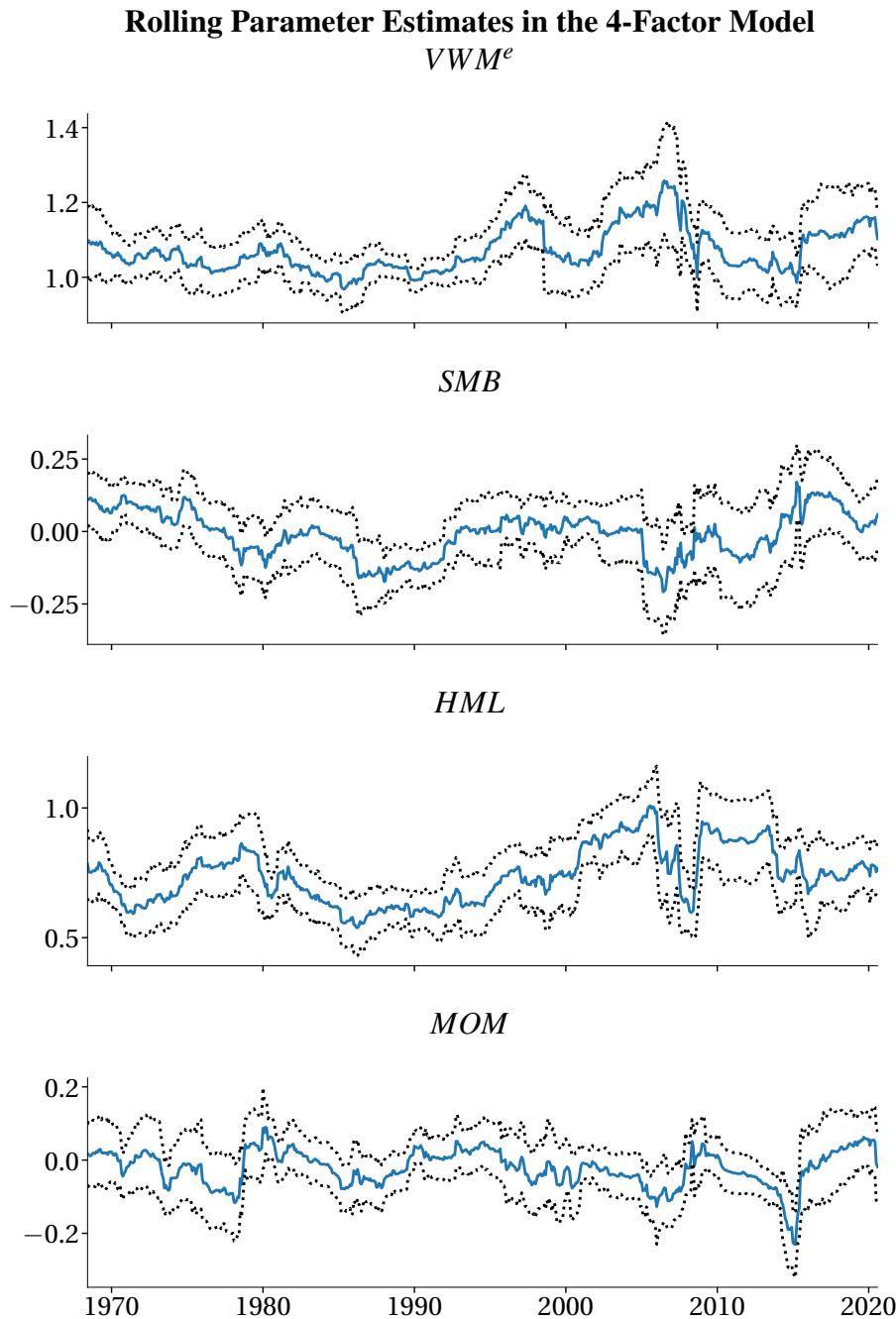


Figure 3.8: 60-month rolling parameter estimates from the model $BH_i^e = \beta_1 + \beta_2 VWM_i^e + \beta_3 SMB_i + \beta_4 HML_i + \beta_5 MOM_i + \varepsilon_i$. Approximate confidence intervals were constructed by scaling the full sample parameter covariance. These rolling estimates indicate that the market loading of the Big-High portfolio varied substantially at the beginning of the samplefixed-length sample and that the loadings on both *SMB* and *HML* may be time-varying.

Recursive Parameter Estimates in the 4-Factor Model

VWM^e

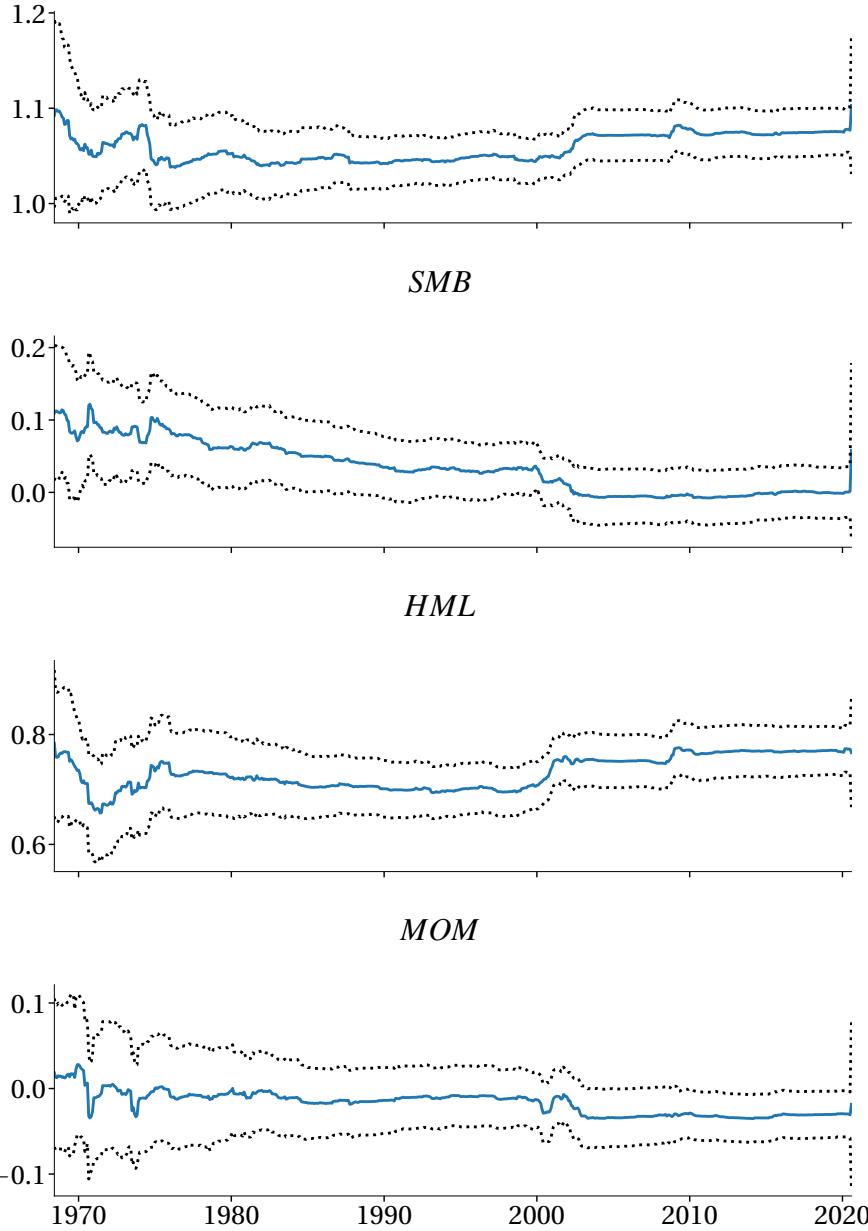


Figure 3.9: Recursive parameter estimates from the model $BH_i^e = \beta_1 + \beta_2 VWM_i^e + \beta_3 SMB_i + \beta_4 HML_i + \beta_5 MOM_i + \varepsilon_i$. Approximate confidence intervals were constructed by scaling the full sample parameter covariance. While less compelling than the rolling window estimates, these recursive estimates indicate that the loading on the market and on HML may not be constant throughout the sample.

3.13.2.3 Normality

Normality may be a concern if the validity of the small-sample assumptions is important. The standard method to test for normality of estimated residuals is the Jarque-Bera (JB) test which is based on two higher order moments (skewness and kurtosis) and tests whether they are consistent with those of a normal distribution. In the normal, the skewness is 0 (it is symmetric) and the kurtosis is 3. Let $\hat{\varepsilon}_i$ be the estimated residuals. Skewness and kurtosis are defined

$$\begin{aligned}\hat{s}k &= \frac{n^{-1} \sum_{i=1}^n \hat{\varepsilon}_i^3}{(\hat{\sigma}^2)^{\frac{3}{2}}} \\ \hat{\kappa} &= \frac{n^{-1} \sum_{i=1}^n \hat{\varepsilon}_i^4}{(\hat{\sigma}^2)^2}\end{aligned}$$

The JB test is computed

$$JB = \frac{n}{6} \left(sk^2 + \frac{1}{4} (\kappa - 3)^2 \right)$$

and is distributed χ_2^2 . If $sk \approx 0$ and $\kappa \approx 3$, then the JB should be small and normality should not be rejected. To use the JB test, compute JB and compare it to C_α where C_α is the critical value from a χ_2^2 . If $JB > C_\alpha$, reject the null of normality.

3.13.2.4 Heteroskedasticity

Heteroskedasticity is a problem if neglected. See section 3.12.4.

3.13.2.5 Influential Observations

Influential observations are those which have a large effect on the estimated parameters. Data, particularly data other than asset price data, often contain errors.²⁴ These errors, whether a measurement problem or a typo, tend to make $\hat{\beta}$ unreliable. One method to assess whether any observation has an undue effect on the sample is to compute the vector of “hat” matrices,

$$h_i = \mathbf{x}_i (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i'$$

This vector (which is the diagonal of $\mathbf{P}_{\mathbf{X}}$) summarizes the influence of each observation on the estimated parameters and is known as the influence function. Ideally, these should be similar and no observation should dominate.

Consider a simple specification where $Y_i = X_i + \varepsilon_i$ where X_i and ε_i are i.i.d. standard normal. In this case the influence function is well behaved. Now suppose one x_i is erroneously increased by 100. In this case, the influence function shows that the contaminated observation (assume it is X_n) has a large impact on the parameter estimates. Figure 3.10 contains four panels. The two left panels show the original data (top) and the data with the error (bottom) while the two right panels contain the influence functions. The influence function for the non-contaminated data is well behaved and each observation has less than 10% influence. In the contaminated data, one observation (the big outlier), has an influence greater than 98%.

²⁴ And even some asset price data, such as TAQ prices.

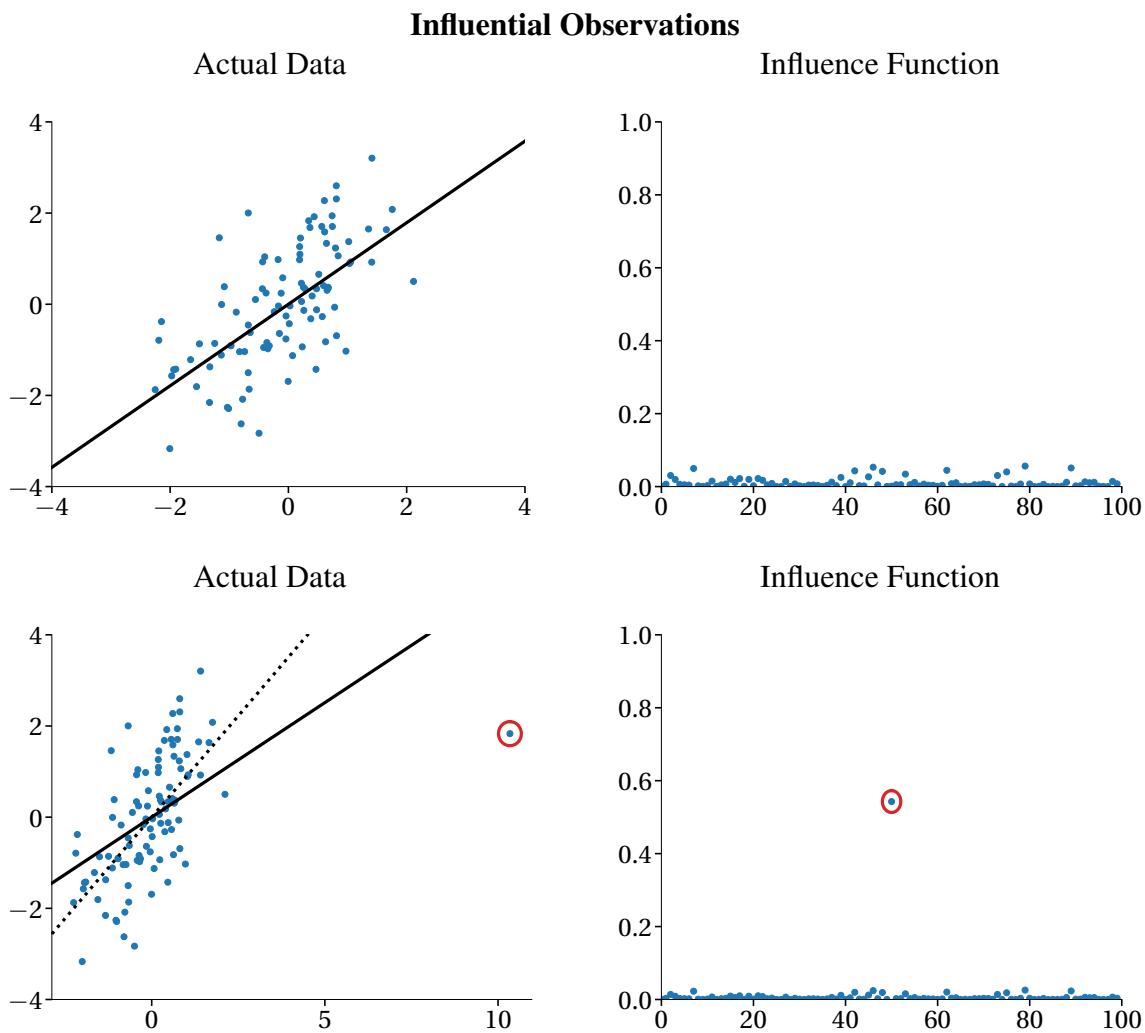


Figure 3.10: The two left panels contain realizations from the data generating process $Y_i = X_i + \varepsilon_i$ where a single X_i has been contaminated (bottom left panel). The two right panels contain the influence functions of the X_i . If all data points were uniformly influential, the distribution of the influence function should be close to uniform (as is the case in the top left panel). In the bottom right panel, it is clear that the entire fit is being driven by a single X_i which has an influence greater than .98.

Plotting the data would have picked up this problem immediately. However, it may be difficult to determine whether an observation is influential when using multiple regressors because the regressors for an observation may be “large” in many dimensions.

3.13.3 Improving estimation in the presence of outliers

Data may contain outliers for many reasons: someone entered an incorrect price on an electronic exchange, a computer glitch multiplied all data by some large constant or a CEO provided an answer out-of-line with other answers due to misunderstanding a survey question. The standard least-squares

estimator is non-robust in the sense that large observations can have a potentially unbounded effect on the estimated parameters. A number of techniques have been developed to produce “robust” regression estimates that use weighted least squares to restrict the influence of any observation.

For clarity of exposition, consider the problem of estimating the mean using data that may be contaminated with a small number of large errors. The usual estimator will be heavily influenced by these outliers, and if outliers occur with any regularity in the data (suppose, for example, 1% of data is contaminated), the effect of outliers can result in an estimator that is biased and in some cases inconsistent. The simplest method to robustly estimate the mean is to use an α -trimmed mean where α represents a quantile of the empirical distribution of the data.

Definition 3.20 (α -Trimmed Mean). The α -quantile trimmed mean is

$$\hat{\mu}_\alpha = \frac{\sum_{i=1}^n Y_i I_{[C_L \leq Y_i \leq C_U]}}{n^*} \quad (3.110)$$

where $n^* = n(1 - \alpha) = \sum_{i=1}^n I_{[-C < Y_i < C]}$ is the number of observations used in the trimmed mean.²⁵

Usually α is chosen to be between .90 and .99. To use an α -trimmed mean estimator, first compute C_L the $\alpha/2$ -quantile and C_U the $1 - \alpha/2$ -quantile of the of y . Using these values, compute the trimmed mean as

A closely related estimator to the trimmed mean is the Winsorized mean. The sole difference between an α -trimmed mean and a Winsorized mean is the method for addressing the outliers. Rather than dropping extreme observations below C_L and C_U , a Winsorized mean truncates the data at these points.

Definition 3.21 (Winsorized mean). Let Y_i^* denote a transformed version of Y_i ,

$$Y_i^* = \max(\min(Y_i, C_U), C_L)$$

where C_L and C_U are the $\alpha/2$ and $1 - \alpha/2$ quantiles of Y . The Winsorized mean is defined

$$\hat{\mu}_W = \frac{\sum_{i=1}^n Y_i^*}{n}. \quad (3.111)$$

While the α -trimmed mean and the Winsorized mean are “robust” to outliers, they are not robust to other assumptions about the data. For example, both mean estimators are biased unless the distribution is symmetric, although “robust” estimators are often employed as an ad-hoc test that results based on the standard mean estimator are not being driven by outliers.

Both of these estimators are in the family of linear estimators (L-estimators). Members of this family can always be written as

$$\hat{\mu}^* = \sum_{i=1}^n w_i Y_i$$

for some set of weights w_i where the data, Y_i , are ordered such that $Y_{j-1} \leq Y_j$ for $j = 2, 3, \dots, N$. This class of estimators obviously includes the sample mean by setting $w_i = \frac{1}{n}$ for all i , and it also includes the median by setting $w_i = 0$ for all i except $w_m = 1$ where $m = (n+1)/2$ (n is odd) or $w_m = w_{m+1} = 1/2$ where $m = n/2$ (n is even). The trimmed mean estimator can be constructed by

²⁵This assumes that $n\alpha$ is an integer. If this is not the case, the second expression is still valid.

setting $w_i = 0$ if $n \leq s$ or $i \geq n - s$ and $w_i = \frac{1}{n-2s}$ otherwise where $s = n\alpha$ is assumed to be an integer. The Winsorized mean sets $w_i = 0$ if $n \leq s$ or $n \geq N - s$, $w_i = \frac{s+1}{n}$ if $n = s + 1$ or $n = n - s - 1$ and $w_i = \frac{1}{n}$ otherwise. Examining the weights between the α -trimmed mean and the Winsorized mean, the primary difference is on the weights w_{k+1} and w_{n-k-1} . In the trimmed mean, the weights on these observations are the same as the weights on the data between these points. In the Winsorized mean estimator, the weights on these observations are $\frac{k+1}{n}$ reflecting the censoring that occurs at these observations.

3.13.3.1 Robust regression-based estimators

Like the mean estimator, the least-squares estimator is not “robust” to outliers. To understand the relationship between L-estimators and linear regression, consider decomposing each observation into its mean and an additive error,

$$\begin{aligned}\hat{\mu}^* &= \sum_{i=1}^n w_i Y_i \\ &= \sum_{i=1}^n w_i (\mu + \varepsilon_i) \\ &= \sum_{i=1}^n w_i \mu + \sum_{i=1}^n w_i \varepsilon_i\end{aligned}$$

A number of properties can be discerned from this decomposition. First, in order for μ^* to be unbiased it must be the case that $\sum_{i=1}^n w_i = 1$ and $\sum_{i=1}^n E[w_i \varepsilon_i] = 0$. All of the linear estimators satisfy the first condition although the second will depend crucially on the distribution of the errors. If the distribution of the errors is symmetric then the Winsorized mean, the α -trimmed mean or even median are unbiased estimators of the mean. However, if the error distribution is not symmetric, then these estimators are likely to be biased. Unlike the usual case where $E[w_i \varepsilon_i] = w_i E[\varepsilon_i]$, the weights are functions of the errors and the expectation of the product of the expectations is not the expectation of the product. Second, weights on the observations (Y_i) are the same as weights on the errors, ε_i . This relationship follows from noticing that if $Y_j \leq Y_{j+1}$, then it must be the case that $\varepsilon_j \leq \varepsilon_{j+1}$.

Robust estimators in linear regression models require a two-step or iterative procedure. The difference between robust mean estimators and robust regression arises since if Y_i has a relationship to a set of explanatory variables \mathbf{x}_i , then orderings based on Y_i will not be the same as orderings based on the residuals, ε_i . For example, consider the simple regression

$$Y_i = \beta X_i + \varepsilon_i.$$

Assuming $\beta > 0$, the largest Y_i are those which correspond either the largest X_i or ε_i . Simple trimming estimators will not only trim large errors but will also trim Y_i that have large values of X_i . The left panels of figure 3.11 illustrate the effects of Winsorization and trimming on the raw data. In both cases, the regression coefficient is asymptotically biased (as indicated by the dotted line) since trimming the raw data results in an error that is correlated with the regressor. For example, observations with the largest X_i values and with positive ε_i more likely to be trimmed. Similarly, observations for

the smallest X_i values and with negative ε_i are more likely to be trimmed. The result of the trimming is that the remaining ε_i are *negatively* correlated with the remaining X_i .

To avoid this issue, a two-step or iterative procedure is needed. The first step is used to produce a preliminary estimate of $\hat{\beta}$. OLS is commonly used in this step although some other weighted least-squares estimator may be used instead. Estimated residuals can be constructed from the preliminary estimate of β ($\hat{\varepsilon}_i = Y_i - \mathbf{x}_i\hat{\beta}$), and the trimming or Winsorizing is done on these preliminary residuals. In the case of α -trimming, observations with the largest errors (in absolute value) are dropped, and the α -trimmed regression is estimated using only the observations with $C_L < \hat{\varepsilon}_i < C_U$.

Winsorized regression also uses the first step regression to estimate $\hat{\varepsilon}$, but, rather than dropping observations, errors larger than C_U are set to $\hat{\varepsilon}_U$ and errors smaller than C_L are set to $\hat{\varepsilon}_L$. Using these modified errors,

$$\hat{\varepsilon}_i^* = \max(\min(\hat{\varepsilon}_i, C_U), C_L)$$

a transformed set of dependent variables is created, $Y_i^* = \mathbf{x}_i\hat{\beta} + \hat{\varepsilon}_i^*$. The Winsorized regression coefficients are then estimated by regressing Y_i^* on \mathbf{x}_i . The correct application of α -trimming and Winsorization are illustrated in the bottom two panels of figure 3.11. In the α -trimming examples, observations marked with an \mathbf{x} were trimmed, and in the Winsorization example, observations marked with a \bullet were reduced from their original value to either C_U or C_L . It should be noted that while both of these estimators are unbiased, this result relies crucially on the symmetry of the errors.

In addition to the two-step procedure illustrated above, an iterative estimator can be defined by starting with some initial estimate of $\hat{\beta}$ denoted $\hat{\beta}^{(1)}$ and then trimming (or Winsorization) the data to estimate a second set of coefficients, $\hat{\beta}^{(2)}$. Using $\hat{\beta}^{(2)}$ and the original data, a different set of estimated residuals can be computed $\hat{\varepsilon}_i = Y_i - \mathbf{x}_i\hat{\beta}^{(2)}$ and trimmed (or Winsorized). Using the new set of trimmed observations, a new set of coefficients, $\hat{\beta}^{(3)}$, can be estimated. This procedure can be repeated until it converges – max $\left| \hat{\beta}^{(i)} - \hat{\beta}^{(i-1)} \right|$.²⁶

Both α -trimmed and Winsorized regression are special cases of a broader class of “robust” regression estimators. Many of these robust regression estimators can be implemented using an iterative procedure known as Iteratively Re-weighted Least Squares (IRWLS) and, unlike trimmed or Winsorized least squares, are guaranteed to converge. For more on these estimators, see Huber (2004) or Rousseeuw and Leroy (2003).

3.13.3.2 Ad-hoc “Robust” Estimators

It is not uncommon to see papers that use Winsorization (or trimming) in the academic finance literature as a check that the findings are not being driven by a small fraction of outlying data. This is usually done by directly Winsorizing the dependent variable and the regressors. While there is no theoretical basis for these ad-hoc estimators, they are a useful tool to ensure that results and parameter estimates are valid for “typical” observations as well as for the full sample. However, if this is the goal, other methods, such as visual inspections of residuals or residuals sorted by explanatory variables, are equally valid and often more useful in detecting problems in a regression.

²⁶These iterative procedures may not converge due to cycles in $\{\hat{\beta}^{(j)}\}$.

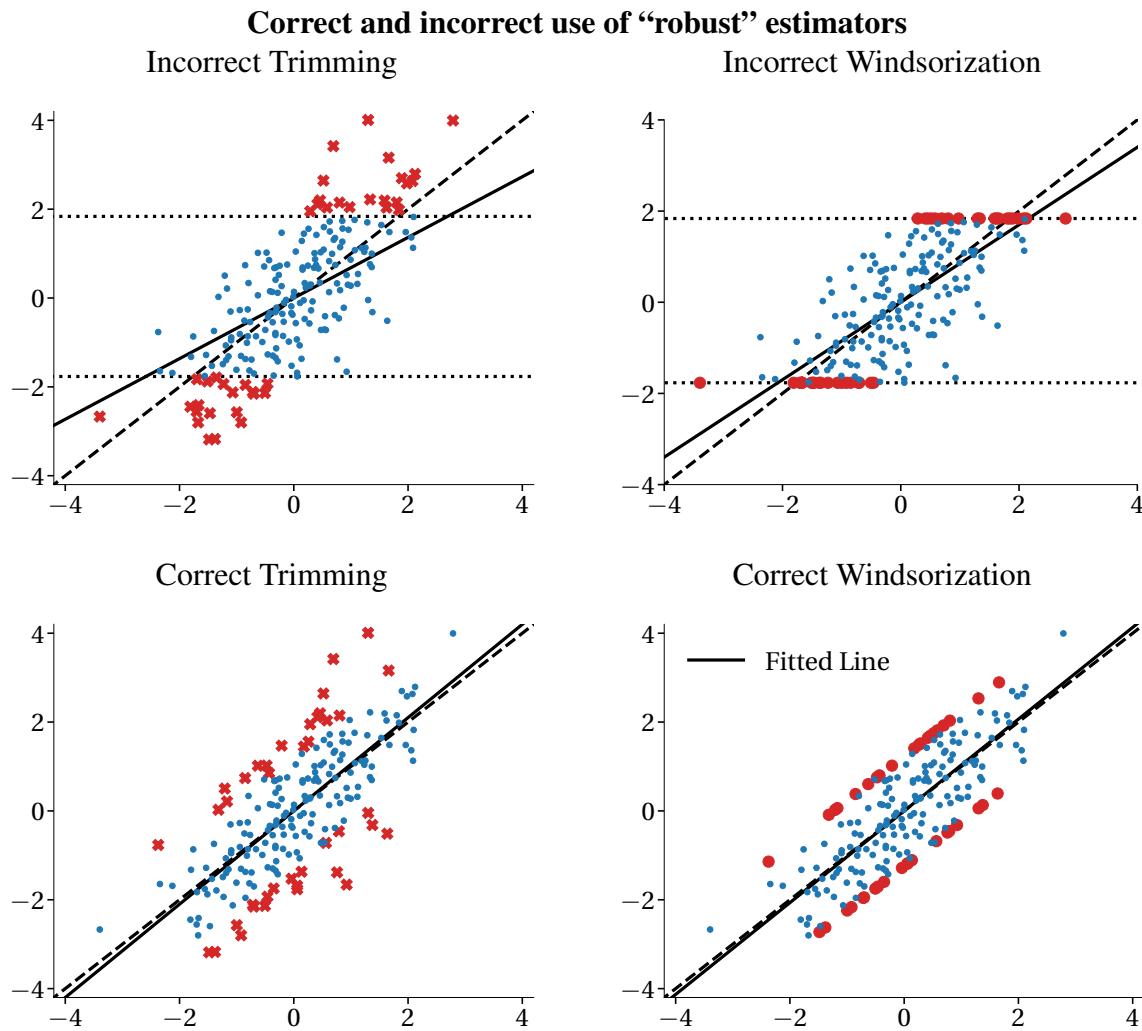


Figure 3.11: These four panels illustrate correct and incorrect α -trimming (left) and Windsorization (right). In both cases, the DGP was $Y_i = X_i + \varepsilon_i$ where X_i and ε_i were independent standard normal random variables. The top panels show incorrect trimming based on the unmodified data, and the bottom panels show correct trimming based on an initial estimate of the slope.

3.13.3.3 Inference on "Robust" Estimators

It may be tempting to use OLS or White heteroskedasticity robust standard errors in "robust" regressions. These regressions (and most L-estimators) appear similar to standard least-squares estimators. However, there is an additional term in the asymptotic covariance of the estimated regression coefficients since the trimming or Windsorization point must be estimated. This term is related to the precision of the trimming point and is closely related to the uncertainty affecting the estimation of a quantile. Fortunately, bootstrapping can be used (under some mild conditions) to estimate the covariance of the regressors.

3.14 Machine Learning

Machine learning approaches to regression, also known as supervised learning, address two key challenges:

- Variable selection when the number of candidate variables is large. In machine learning, variables are often called features, and the collection of all features is called the feature space. Most machine learning algorithms are capable of modeling data sets where the number of variables exceeds the number of observations available.
- Optimizing model parameters to perform well in out-of-sample prediction. In most applications, this optimization makes an explicit trade-off between bias and variance, and most ML approaches to regression use biased estimators that have lower parameter variance than vanilla OLS. This reduction in variance, especially for parameters that have a small effect relative to their uncertainty, improves out-of-sample prediction at the cost of some bias.

ML approaches achieve these goals using cross-validation to select models and parameter values that perform well both in- and out-of-sample. These alternative approaches generally provide methods to jointly select relevant variables and estimate parameters. Some methods make use of bootstrapping to improve the reliability of the models in out-of-sample data. Ultimately these approaches all produce a standard linear regression model where the coefficients are not usually estimated using standard OLS. The most useful strategies tend to introduce a limited amount of bias by *shrinking* regression coefficients toward 0 to mitigate the cost of parameter uncertainty.

3.14.1 Best Subset Regression

Best Subset Regression is the simplest method to construct a model given a set of predictors. Suppose you have p candidate variables $X_{1,i}, \dots, X_{p,i}$. Best Subset Regression finds the combination of variables in this set that optimizes the model's fit according to some criteria, for example, the cross-validated SSE or BIC. Best Subset Regression begins by finding the model that produces the smallest in-sample SSE, or equivalently the largest R^2 , using k of the p variables. Let this model be denoted \mathcal{M}_k . This step involves fitting $\binom{p}{k}$ distinct models. The best model is selected for each possible value of $k = 1, 2, \dots, p$. The initial inputs are a set of $p + 1$ distinct models $\mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_p$ where \mathcal{M}_0 is a model that contains no predictors. The Best Subset Regression is chosen by comparing the performance of these $p + 1$ models using some criterion, for example, the cross-validated SSE, and selecting the model that performs the best. There are two important issues with Best Subset Regression. First, it can only be used when the set of candidate predictors p is moderate (≤ 30) since there are $2^p - 1$ distinct models that must be estimated. Second, the coefficients of the best model are estimated by OLS. OLS estimates always overfit the sample used to estimate the parameters, and the in-sample overfitting reduces the out-of-sample performance of the models.

Algorithm 3.11 (Best Subset Regression).

1. For $k \in \{0, 1, \dots, p\}$ estimate each of the $\binom{p}{k}$ distinct models containing k variables, saving the model that produces the smallest SSE as \mathcal{M}_j , $j = 0, \dots, p$.

2. Select the Best Subset Regression as the model from the set $\{\mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_p\}$ that minimizes some criterion such as the cross-validated SSE.

3.14.2 Forward, Backward, and Hybrid Stepwise Regression

Best Subset Regression cannot be used when p is large. Stepwise model building is an alternative that builds the models $\mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_p$ sequentially. Forward stepwise regression begins with no variables selected. Each of the excluded variables, p in total, are tried one at a time, and the regressor that produces the best fit is retained in \mathcal{M}_1 . The second model, \mathcal{M}_2 , is then selected by adding each of the $p - 1$ variables that were not included in \mathcal{M}_1 and is defined as the model that produces the best in-sample fit. This process is repeated so that \mathcal{M}_{j+1} adds one of the $p - j$ variables to \mathcal{M}_j that were not included in \mathcal{M}_j . The output of the first step is a set of $p + 1$ models $\mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_p$ where larger models always nest smaller models. The final model is selected from the set of candidate models by optimizing some criterion such as the cross-validated SSE.

Algorithm 3.12 (Forward Stepwise Regression).

1. Begin with the empty model, \mathcal{M}_0 .
2. For $j \in \{0, \dots, p - 1\}$, construct model \mathcal{M}_{j+1} as the model that minimizes the SSE by adding each of the $p - j$ variables to the variables included in model \mathcal{M}_j .
3. Select the Forward Stepwise Regression as the model from the set $\{\mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_p\}$ that minimizes some criterion such as the cross-validated SSE.

Backward stepwise regression operates in the opposite direction. Begin with the model that contains all variables \mathcal{M}_p . The next smaller model, \mathcal{M}_{p-1} is defined as the model that minimizes the SSE considering each of the p models that drops a single variable from \mathcal{M}_p . This process continues where \mathcal{M}_j is defined as the model that maximizes the in-sample fit using j of the $j + 1$ variables included in \mathcal{M}_{j+1} . Like forward stepwise regression, backward stepwise regression produces a set of $p + 1$ models $\mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_p$. The best model is then selected from this set of candidate models by optimizing some criterion function.

Algorithm 3.13 (Backward Stepwise Regression).

1. Begin with the complete model, \mathcal{M}_p .
2. For $j \in \{p - 1, p - 2, \dots, 0\}$, construct model \mathcal{M}_j as the model that minimizes the SSE by removing each of the j variables, one at a time, of the variables included in model \mathcal{M}_{j+1} .
3. Select the Backward Stepwise Regression as the model from the set $\{\mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_p\}$ that minimizes some criterion such as the cross-validated SSE.

Hybrid approaches combine the two. For example, suppose forward stepwise regression is used to select \mathcal{M}_k where $k < p$. Backward stepwise regression can be used on the k included regressors in \mathcal{M}_k to produce a new sequence of models \mathcal{M}_j^k for $j = k - 1, k - 2, \dots, 1$. This sequence may be distinct from what forward or backward stepwise regression would arrive at alone. The hybrid approach generally produces a larger set of candidate models while remaining computationally tractable as long as the number of direction switches is small. This larger set of candidate models has an increased chance

of including the Best Subset Regression than either forward or backward stepwise regression alone. The primary challenge of the hybrid approach is determining the number of direction reversals to use, although, in practice, this is often dictated by the computational time available. Like both forward and backward stepwise regression, the final model is selected from the enlarged pool of candidate models by optimizing some criteria.

3.14.3 Ridge Regression

Ridge regression differs from best subset and stepwise regression in two ways: it does not select variables, and coefficients are not estimated using standard OLS.

Definition 3.22 (Ridge Regression).

The ridge regression estimator with tuning parameter ω is defined as the solution to

$$\arg \min_{\beta} (\mathbf{y} - \mathbf{X}\beta) (\mathbf{y} - \mathbf{X}\beta) \text{ subject to } \sum_{j=1}^k \beta_j^2 \leq \omega. \quad (3.112)$$

This constrained problem is equivalent to the unconstrained problem

$$\arg \min_{\beta} (\mathbf{y} - \mathbf{X}\beta) (\mathbf{y} - \mathbf{X}\beta) + \lambda \sum_{j=1}^k \beta_j^2 \quad (3.113)$$

where ω and λ take different values and have an inverse relationship (i.e., large values of ω correspond to small values of λ). The solution to this optimization problem is

$$\hat{\beta}^{\text{Ridge}} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k)^{-1} \mathbf{X}'\mathbf{y} \quad (3.114)$$

where k is the number of regressors included in the model.

Recall that the OLS estimator is $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$. The effect of the ridge penalty is simple to deduce from eq. (3.114) since $\lambda > 0$. The term $\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k$ must always be larger, in a matrix sense, than $\mathbf{X}'\mathbf{X}$ since $\lambda \mathbf{I}_k$ is a diagonal matrix with positive values along its diagonal. It must then be the case that $(\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_k)^{-1}$ is *smaller* than $\mathbf{X}'\mathbf{X}$, again in a matrix sense, and so the ridge coefficient estimates $\hat{\beta}^{\text{Ridge}}$ are always closer to 0 than the OLS estimates $\hat{\beta}$. Ridge regression is known as a *shrinkage* estimator since the parameter estimates pull the parameters towards the shrinkage target of 0. In practice shrinkage introduces some bias in the coefficient but reduces their variance, and ridge regression often outperforms OLS in out-of-sample applications.

Ridge regression depends on a single tuning parameter, λ , which controls how bias and variance are traded off. The optimal value is determined by trying several different values and selecting the value λ^* that produces the smallest cross-validated SSE. Note that ridge regression does not provide any guidance as to which variables to include in the model, and so some form of model selection is usually needed. The optimal choice of λ depends on the number of regressors included in the model, and so it must be re-optimized in each distinct model. There are many variants of ridge regression that change the penalty structure. For example, one variant allows the shrinkage to be applied to only a subset of the included variables. This penalization structure can be useful if some variables are strong predictors, while others are less useful. This penalty structure can be further generalized to

apply different amounts of shrinkage to distinct groups of regressors or even to impose cross-regressor shrinkage where the total magnitude of a set of the regressors in the model is affected.²⁷

3.14.4 LASSO, Forward Stagewise Regression, and LARS

LASSO (least absolute shrinkage and selection operator), Forward Stagewise Regression, and LARS (Least Angle Regression) are relatively new methods that embed both variable selection and shrinkage into a unified approach (Tibshirani, 1996; Efron, Hastie, Johnstone, and Tibshirani, 2004). LASSO is similar to ridge regression and can be written as a constrained least square problem.

Definition 3.23 (LASSO). The LASSO estimator with tuning parameter ω is defined as the solution to

$$\arg \min_{\beta} (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \text{ subject to } \sum_{j=1}^k |\beta_j| < \omega \quad (3.115)$$

The key difference is that the constraint is on the sum of the *absolute value* of the coefficients and not their squared values. The LASSO estimator adds an additional constraint to the least-squares problem that limits the magnitude of regression coefficients that produces an interpretable model. Regressors that have little explanatory power will have coefficients exactly equal to 0 (and hence are excluded). This means that LASSO both estimates parameters and selects variables – any variable with a coefficient that is exactly 0 is effectively removed from the model.

The LASSO constrained minimization problem is dual to a penalized least-squares problem,

$$\arg \min_{\beta} (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + \lambda \sum_{j=1}^k |\beta_j| \quad (3.116)$$

where ω and λ have an inverse relationship. While LASSO has a closed form solution for any value of λ , the estimator is not simple to describe in a single equation.

Forward Stagewise Regression is closely related to LASSO and illustrates the fundamental principle used in variable selection. Estimation begins with a model that contains no regressors. The algorithm then uses an iterative method to build the regression in small steps by expanding the regression coefficients (small enough that the coefficient expansions should be virtually continuous).

²⁷The complete formulation of a ridge regression is

$$\arg \min_{\beta} (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + (\beta - \beta_0)' \Lambda (\beta - \beta_0)$$

where β_0 is the shrinkage target and Λ is a positive definite matrix that controls the amount of shrinkage. This form nests the classic specification when $\Lambda = \lambda \mathbf{I}_k$ and $\beta_0 = 0$. If Λ is not diagonal, then the estimator will apply cross-variable penalties. The solution to the general problem is

$$\hat{\beta}^{\text{Ridge}} = (\mathbf{X}'\mathbf{X} + \Lambda)^{-1} (\mathbf{X}'\mathbf{y} + \Lambda\beta_0).$$

This shows that the OLS solution is recovered when $\Lambda = \mathbf{0}$. If Λ is very large, then $\hat{\beta}^{\text{Ridge}} \approx \Lambda^{-1} \Lambda \beta_0 = \beta_0$ and the estimate depends only on the shrinkage target β_0 .

Algorithm 3.14 (Forward Stagewise Regression). *The Forward Stagewise Regression (FSR) estimator is defined as the sample paths of $\hat{\beta}$ defined by*

1. Begin with $\hat{\beta}^{(0)} = 0$, and errors $\varepsilon^{(0)} = \mathbf{y}$
2. Compute the correlations of the residual at iteration i with the regressors, $\mathbf{c}^{(i)} = \text{Corr} [\mathbf{X}, \varepsilon^{(i)}]$
3. Define j to be the index of the largest element of $|\mathbf{c}^{(i)}|$ (the absolute value of the correlations), and update the coefficients where $\hat{\beta}_j^{(i+1)} = \hat{\beta}_j^{(i)} + \eta \cdot \text{sign}(c_j)$ and $\hat{\beta}_l^{(i+1)} = \hat{\beta}_l^{(i)}$ for $l \neq j$ where η is a small number (should be much smaller than c_j).²⁸
4. Compute $\varepsilon^{(i+1)} = \mathbf{y} - \mathbf{X}\hat{\beta}^{(i+1)}$
5. Repeat steps 2 – 4 until all correlations are 0 (if $\varepsilon^{(i)} = \mathbf{0}$ than all correlations are 0 by definition).

The coefficients of FSR are determined by taking a small step in the direction of the highest correlation between the regressors and the current error, and so the algorithm will always take a step in the direction of the regressor that has the most (local) explanatory power over the regressand. The final stage FSR coefficients will be equal to the OLS estimates as long as the number of regressors under consideration is smaller than the number of observations. The LASSO estimate is usually computed using the LARS algorithm, which simplifies FSR by finding the exact step size needed before the next variable enters the regression.

Algorithm 3.15 (Least Angle Regression). *The Least Angle Regression (LARS) estimator is defined as the sample paths of $\hat{\beta}$ defined by:*

1. Begin with $\hat{\beta}^{(0)} = 0$, and errors $\varepsilon^{(0)} = \tilde{\mathbf{y}}$ where

$$\tilde{\mathbf{y}} = \frac{\mathbf{y} - \bar{y}}{\hat{\sigma}_y} \quad (3.117)$$

and

$$\tilde{\mathbf{x}}_i = \frac{\mathbf{x}_i - \bar{\mathbf{x}}_i}{\hat{\sigma}_x} \quad (3.118)$$

are studentized versions of the original data.²⁹

2. Compute the correlations of the residual at state i with the regressors, $\mathbf{c}^{(i)} = \text{Corr} [\tilde{\mathbf{X}}^{(i)}, \varepsilon^{(i)}]$ and define j to be the index of the largest element of $|\mathbf{c}^{(i)}|$ (the absolute value of the correlations).
3. Define the active set of regressors $\tilde{\mathbf{X}}^{(1)} = \tilde{\mathbf{x}}_j$.

²⁸ η should be larger than some small value to ensure the algorithm completes in finitely many steps, but should always be weakly smaller than $|c_j|$.

²⁹ LARS can be implemented on non-studentized data by replacing correlation with $\mathbf{c}^{(i)} = \mathbf{X}^{(i)'} \varepsilon^{(i)}$.

4. Move $\hat{\beta}^{(1)} = \hat{\beta}_j$ towards the least squares estimate of regressing $\varepsilon^{(0)}$ on $\tilde{\mathbf{X}}^{(1)}$ until the correlation between $\varepsilon^{(1)} = \tilde{\mathbf{y}} - \tilde{\mathbf{X}}^{(1)}\hat{\beta}^{(1)}$ and some other $\tilde{\mathbf{x}}_k$ is equal to the correlation between $\varepsilon^{(1)}$ and $\tilde{\mathbf{x}}_j$.
5. Add $\tilde{\mathbf{x}}_k$ to the active set of regressors so $\tilde{\mathbf{X}}^{(2)} = [\tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_k]$.
6. Move $\hat{\beta}^{(2)} = [\hat{\beta}_j \hat{\beta}_k]$ towards the least squares estimate of regressing $\varepsilon^{(1)}$ on $\tilde{\mathbf{X}}^{(2)}$ until the correlation between $\varepsilon^{(2)} = \tilde{\mathbf{y}} - \tilde{\mathbf{X}}^{(2)}\hat{\beta}^{(2)}$ and some other $\tilde{\mathbf{x}}_l$ is equal to the correlation between $\varepsilon^{(2)}$ and $\tilde{\mathbf{x}}^{(2)}$.
7. Repeat steps 5 – 6 by adding regressors to the active set until all regressors have been added or n steps have been taken, whichever occurs first.

The algorithm of LARS describes the statistical justification for the procedure – variables are added as soon as they have the largest correlation. Once the active set contains two or more regressors, the maximum correlation between the error and all regressors will be the same since regression coefficients are expanded in a manner that keeps the correlation identical between the error and any regressors in the active set. Efron, Hastie, et al. (2004) proposes a new algorithm that allows the entire path of LASSO, FSR, and LARS estimates to be quickly computed in models that contain a large number of candidate regressors. LASSO differs from LARS in one technical aspect, although they are very similar in practice.

These models are deeply related as shown Efron, Hastie, et al. (2004) and Hastie et al. (2007). All three can be used for model selection once a stopping rule (FSR, LARS) or the penalty (λ , LASSO) has been selected. k -fold cross-validation is commonly used to choose these values. Note that the usual standard OLS errors and t -stats are no longer correct since these estimators are constrained versions of least squares. Tibshirani (1996) proposes a bootstrap method that can be used to compute standard errors and make inference on LASSO estimators.³⁰

Figure 3.12 illustrates how ridge regression and LASSO estimate parameters. Both show the OLS estimate $\hat{\beta}$ surrounded by ellipsoids the trace iso-SSE curves – that is, values of β_1 and β_2 that produce the same regression fit. The estimators are defined as the point where the smallest SSE is just tangent to the constraint. The ridge regression shrinks the estimate towards zero in a non-uniform way. This happens since the regressors are correlated. Ridge regression produces an estimate where both coefficients are non-zero. LASSO, on the other hand, estimates β_2 to be exactly. This happens since non-zero β_1 provides a larger reduction in the SSE than β_2 , at least near the point $(0, 0)$. In general, ridge regression will never estimate any coefficients to be exactly 0 except when the OLS coefficient is exactly 0. LASSO frequently estimates coefficients to be zero since the cost of adding a small amount of a coefficient near zero is linear in β while the gain in terms of the SSE is quadratic in β (i.e., $\propto \beta^2$).

Figure 3.13 shows that paths of both the ridge regression and LASSO estimators are the restriction parameter ω is reduced. The model estimated regresses the return on the Big-High portfolio on the four factors, VWM^e , SMB , HML , and MOM . The paths begin with $\omega = 0$. As the constraint is relaxed, the parameters converge towards the OLS estimates, which limit cases as ω increases. There

³⁰The standard errors subsequent to a selection procedure using GtS, StG, or IC are also not correct since tests have been repeated. In this regard, the bootstrap procedure should be more accurate since it accounts for the variation due to the selection, something not usually done in traditional model selection procedures.

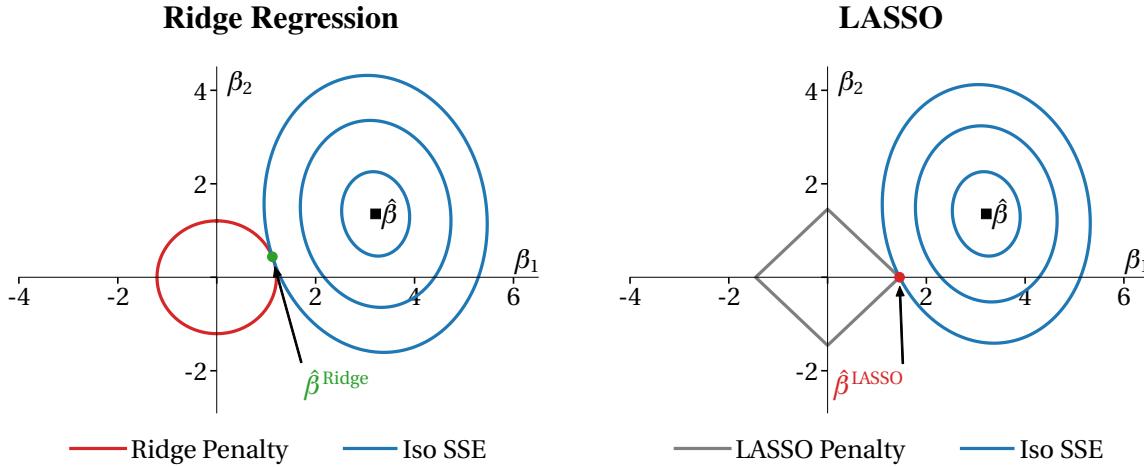


Figure 3.12: The left panel shows the ridge regression restriction for a specific value of ω along with three lines that trace combinations of β_1 and β_2 that produce the same model SSE. The ridge estimate is defined as the point where the SSE is just tangent to a restriction. The right shows the LASSO constraint along with the iso-SSE curves for the same data generating process.

is one clear distinction between the two paths. The paths from ridge regression evolve smoothly as ω increases. All coefficients except *SMB* are different from zero once $\omega > 1/8$. The LASSO paths have a distinct kinked shape. These kinks are points where the correlation between one excluded regressor and the included regressor(s) equalize so that the active set of regressors increases. The market is the strongest predictor, followed by the value factor. Momentum enters the model for small values of the penalty parameter, and size has a non-zero coefficient only at the OLS estimate (and then very small). The dashed line in each plot indicates that optimal choice ω^* selected using 5-fold cross-validation. The cross-validated penalty parameter suggests that little shrinkage is needed. This occurs since the sample size is large enough that parameters, even small values, are precisely estimated.

3.14.5 Regression Trees and their Refinements

Regression Trees build models using only dummy variables. Constructing a regression tree begins by splitting the data into two groups using the values in regressors as possible split values. The model is constructed by splitting the observations into two groups using on all possible values of each regressor. The split that minimizes the SSE is retained, and the two groups are called leaves. The algorithm is then rerun on each leaf again, splitting on all possible values in each of the variables included in the model. This process of splitting into two leaves continues until either the homogeneity in the group as measured by the within-group MSE is sufficiently low, or the number of observations in a leaf falls below some prespecified value.

Figure 3.14 shows the first three levels of a model for the returns on the Big-High portfolio on the four factor portfolios. Splitting the data first on the market produced the largest gains, and the optimal split value was very near zero. The two leaves were then split according to the market into four groups corresponding to very low market returns (≤ -7.17), negative market returns ($-7.17 < VMW \leq -0.81$), positive market returns ($-0.81 < VMW \leq 3.78$), and very high market returns

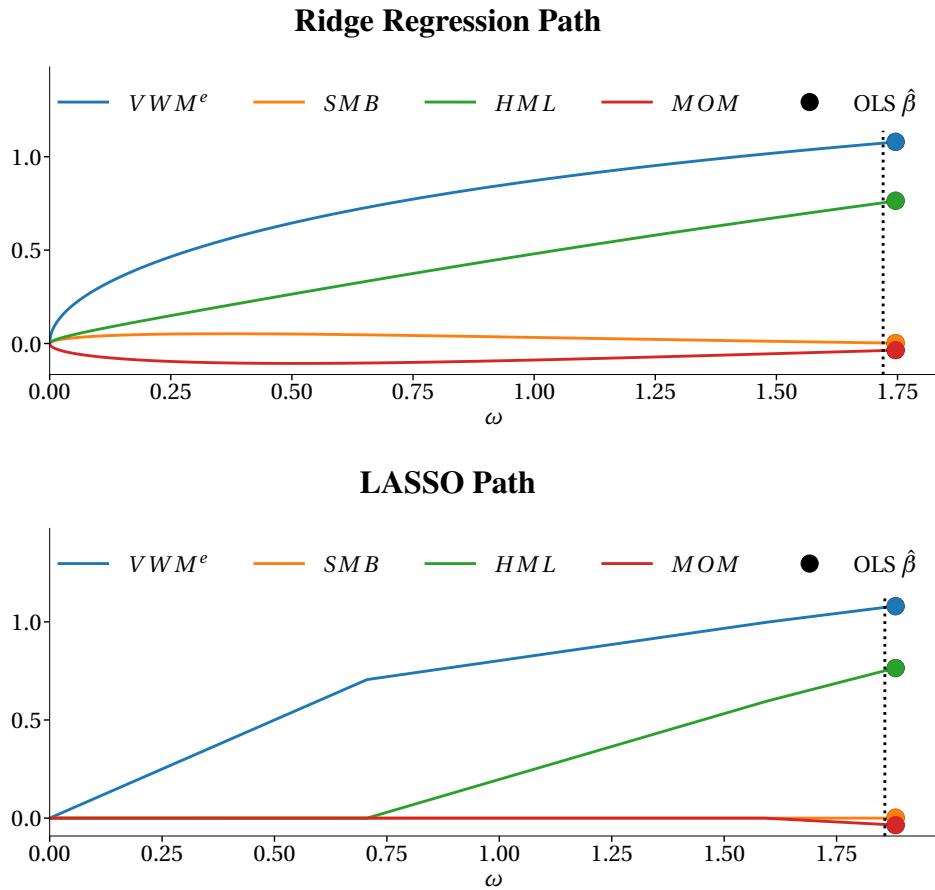


Figure 3.13: The top panel shows the path of the ridge regression estimates from the four factor model $BH^e = \beta_1 + \beta_2 VWM_i^e + \beta_3 SMB_i + \beta_4 HML_i + \beta_5 MOM_i + \varepsilon_i$. The penalty parameter ω is increased from zero to the value that produces the OLS estimate. The bottom panel contains the path of the LASSO estimates as the restriction is decreased. The kinks indicate points where a parameter switches from being exactly zero to a non-zero value.

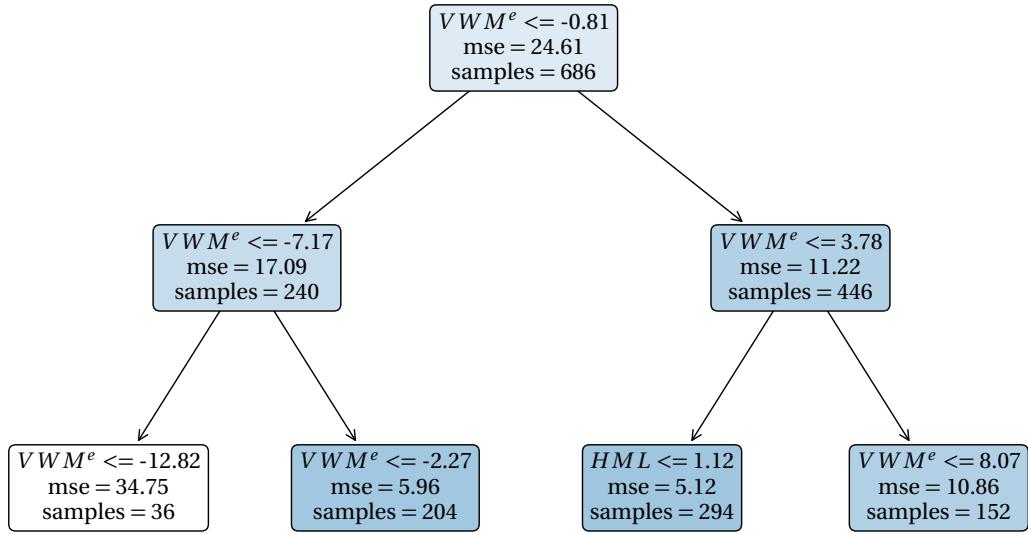


Figure 3.14: A regression tree where the left-hand-side variable is the return on the Big-High portfolio and the model is built using the four factors: VWM^e , SMB , HML , and MOM . The first and second splits used the market portfolio to bin the returns into four regions ranging from very low to very high. The final level splits used different variables so that the terminal leaves depend on both the market and the size factor.

($> 3.78\%$). If the tree was stopped at this node, the regression selected would be

$$BH^e = \beta_1 I_{[VWM_t^e \leq -7.17]} + \beta_2 I_{[-7.17 < VWM_t^e \leq -0.81]} + \beta_3 I_{[-0.81 < VWM_t^e \leq 3.78]} + \beta_4 I_{[VWM_t^e > 3.78]} + \varepsilon_i$$

The estimates of the parameters are simply the within-group means. The final level further splits the data into eight leaves (not shown). Three of the final level splits used the market return to split the negative returns further and to define an extreme positive return leaf. The other split preferred to use value. This final regression model would have eight terms constructed using combinations of restrictions on the return on the market factor and the return of the value factor.

Regression trees have step-function like behavior and frequently are not well suited to analyzing continuous-valued variables using continuously values regressors. While plain regression threes should usually be avoided, four refinements, pruning, Random Forests, bagging, and boosting all produce improvements in regression-tree models. Figure 3.15 compares a 2-level tree with OLS when modeling the return of the Big-High portfolio using the excess market return. The tree approximates the regression line as a step function. While this fit is not a terrible description of the data near 0, there are obvious deficiencies in the tails.

3.14.5.1 Improving Regression Trees

Three techniques are commonly used to improve regression trees: pruning, bagging, boosting, and Random Forests. Pruning a tree removes nodes that make a negligible improvement to the in-sample fit and often decrease out-of-sample fit. Pruning is implemented by optimizing the modified objective function

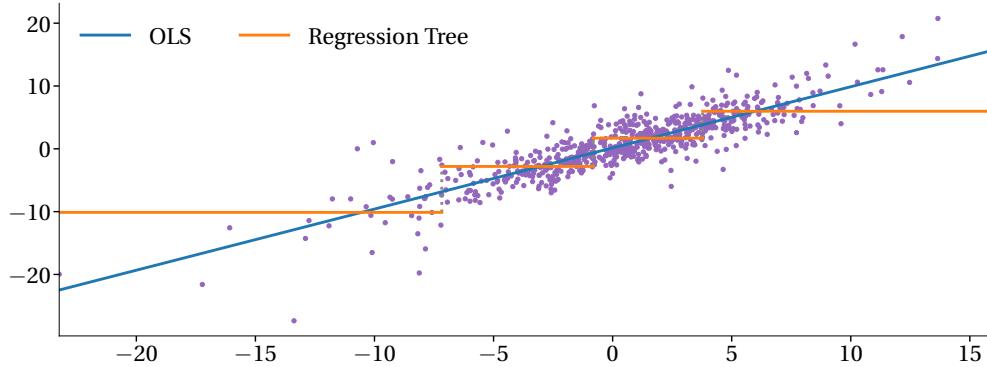


Figure 3.15: The regression tree implied by the first two splits and the OLS fit of the excess returns on the Big-High portfolio on the market.

$$\sum_{i=1}^n (Y_i - \hat{f}(\mathbf{x}_i))^2 + \alpha |T|$$

where $\hat{f}(\mathbf{x}_i)$ is the predicted value for a given tree and $|T|$ is the number of terminal nodes in the tree. Pruning starts with a large tree with T_0 nodes that is only terminated when either the number of nodes hits some threshold, the maximum number of levels is reached, or a SSE-based stopping criterion is met. For values of α on a grid of plausible values $\{\alpha_1 < \alpha_2 < \dots < \alpha_q\}$ the tree that minimizes the modified objective function is selected. The preferred value of $\hat{\alpha}$ is chosen from this grid using k -fold cross-validation. Finally, the pruned tree is estimated by minimizing the modified objective function using $\hat{\alpha}$ on the original sample.

Bagging makes use of B bootstrap samples to the parameters of multiple trees. Each tree can then be used to generate predictions for any value of the regressor \mathbf{x} . These predictions are then be averaged to produce the bagged forecast. Note that each tree may have both a different structure and parameter values. While the forecasts will tend to be similar, they are not perfectly correlated, and the average forecast has a lower variance than any of the individual forecasts.

Algorithm 3.16 (Bagging Regression Trees). *A bagged prediction from a regression tree is constructed following:*

1. For $i = 1, 2, \dots, B$ generate a bootstrap sample from (Y_i, \mathbf{x}_i) and fit a regression tree to the bootstrapped sample.
2. Using the B trees, construct the forecast as

$$\hat{f}(\mathbf{x}) = 1/B \sum_{i=1}^B \hat{f}_i(\mathbf{x})$$

where $f_i(\mathbf{x})$ is the prediction from the tree estimated using bootstrap sample i .

Random Forests make use of randomization by selecting a subset of the available regressors when estimating a tree. When the number of regressors p is large, most trees will tend to have a very

similar structure even when fit to bootstrapped samples. This structure arises since strong predictors will always be selected in the first levels of the tree. The Random Forest solution is to estimate a tree using a bootstrap sample that also random selects $\approx \sqrt{p}$ regressors. This fitting of trees to random subsamples of the data is repeated many times, and the Random Forecast forecast is the average of forecasts of these models. The distinct trees tend to have low correlation, which translates into large gains from averaging.

Algorithm 3.17 (Random Forests). *A Random Forest of regression trees is constructed following:*

1. *For $i = 1, 2, \dots, B$ generate a bootstrap sample of the data with a random subset of $k \approx \sqrt{p}$ regressors and fit a regression tree using the selected subset of the regressors.*
2. *Using the B trees, construct the forecast as*

$$\hat{f}(\mathbf{x}) = 1/B \sum_{i=1}^B \tilde{f}_i(\mathbf{x})$$

where \tilde{f}_i is the prediction using random regressor subset i .

Note that a Random Forest is identical to a bagged regression tree when $k = p$ regressors are used to build each tree.

Boosting also fits multiple trees, only sequentially to the same data. A boosted tree begins by fitting a small tree with d nodes to the data and computing the residuals. It then fits a new tree to the residuals. This is repeated many times. The trees are then combined using a tuning parameter λ as

$$\hat{f}(\mathbf{x}) = 1/B \sum_{i=1}^B \lambda \ddot{f}_i(\mathbf{x})$$

where \ddot{f}_1 is the tree fit to the original data and \ddot{f}_j , $j \geq 2$ is the prediction from the tree estimated using the residuals of the form

$$\hat{\epsilon}_{i,j} = \hat{\epsilon}_{i,j-1} - \lambda f_{j-1}(\mathbf{x}_i)$$

where $\hat{\epsilon}_{i,0} = Y_i$.

Algorithm 3.18 (Bagging Regression Trees). *Begin with $\hat{\epsilon}_{i,0} = \tilde{Y}_i$ where \tilde{Y}_i is the standardized version of Y_i . For $j = 1, \dots, B$:*

1. *Fit a regression tree using $(\hat{\epsilon}_{i,j-1}, \mathbf{x}_i)$ with d splits and generate $\hat{\epsilon}_{i,j} = \hat{\epsilon}_{i,j-1} - \lambda \ddot{f}_j(\mathbf{x}_i)$ where \ddot{f}_j is the tree fit in iteration j .*
2. *Produce the boosted forecast as*

$$\hat{f}(\mathbf{x}) = 1/B \sum_{i=1}^B \lambda \ddot{f}_i(\mathbf{x})$$

Boosting makes uses of three tuning parameters, λ , d , and B . λ is usually set to some small value in the range $(0.001, 0.10)$. Small values of λ slow the learning since much of the forecast is down-weighted. d , the number of terminal nodes in a tree, is also set to some small number, often

1. d determines the maximum number of interactions allowed between the regressors when building the dummy-variable representation of a regression tree. Finally, B is usually set to some large value, often in the range of 1,000 – 10,000. These three parameters all interact and are substitutes – increases in one should usually be matched by decreases in the others when building optimal predictions. All three can be selected using a grid of values and k -fold cross-validation.

3.15 Projection

Least squares has one further justification: it is the best linear predictor of a dependent variable where best is interpreted to mean that it minimizes the mean square error (MSE). Suppose $f(\mathbf{x})$ is a function of only \mathbf{x} and not Y . Mean squared error is defined

$$\text{E}[(Y - f(\mathbf{x}))^2].$$

Assuming that it is permissible to differentiate under the expectations operator, the solution is

$$\text{E}[Y - f(\mathbf{x})] = 0,$$

and, using the law of iterated expectations,

$$f(\mathbf{x}) = \text{E}[y|\mathbf{x}].$$

If $f(\mathbf{x})$ is restricted to include only linear functions of \mathbf{x} then the problem simplifies to choosing β to minimize the MSE,

$$\text{E}[(Y - \mathbf{x}\beta)^2]$$

and differentiating under the expectations (again, when possible),

$$\text{E}[\mathbf{x}'(Y - \mathbf{x}\beta)] = \mathbf{0}$$

and $\hat{\beta} = \text{E}[\mathbf{x}'\mathbf{x}]^{-1}\text{E}[\mathbf{x}'\mathbf{y}]$. In the case where \mathbf{x} contains a constant, this allows the best linear predictor to be expressed in terms of the covariance matrix of y and $\tilde{\mathbf{x}}$ where the \sim indicates the constant has been excluded (i.e., $\mathbf{x} = [1 \tilde{\mathbf{x}}]$), and so

$$\hat{\beta} = \Sigma_{\mathbf{xx}}^{-1}\Sigma_{\mathbf{xy}}$$

where the covariance matrix of $[Y \tilde{\mathbf{x}}]$ can be partitioned

$$\text{Cov}([Y \tilde{\mathbf{x}}]) = \begin{bmatrix} \Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}} \\ \Sigma'_{\mathbf{xy}} & \Sigma_{yy} \end{bmatrix}$$

Recall from assumptions 3.7 that $\{\mathbf{x}_i, \varepsilon_i\}$ is a stationary and ergodic sequence and from assumption 3.8 that it has finite second moments and is of full rank. These two assumptions are sufficient to justify the OLS estimator as the best linear predictor of Y . Further, the OLS estimator can be used to make predictions for out of sample data. Suppose Y_{n+1} was an out-of-sample data point. Using the OLS procedure, the best predictor of Y_{n+1} (again, in the MSE sense), denoted \hat{Y}_{n+1} is $\mathbf{x}_{n+1}\hat{\beta}$.

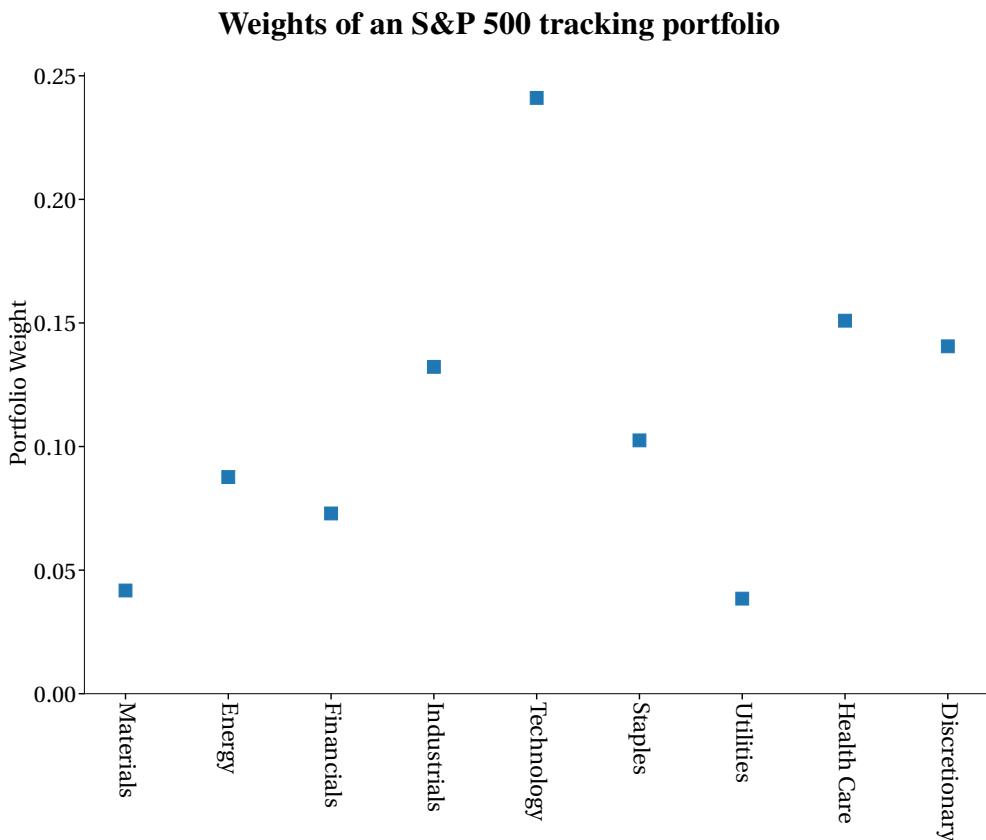


Figure 3.16: Plot of the optimal tracking portfolio weights. The optimal tracking portfolio is long all asset and no weight is greater than 25%.

3.15.1 Tracking Error Minimization

Consider the problem of setting up a portfolio that would generate returns as close as possible to the return on some index, for example, the FTSE 100. One option would be to buy the entire portfolio and perfectly replicate the portfolio. For other indices, such as the Wilshire 5000, which consists of many small and illiquid stocks, complete replication is impossible, and a tracking portfolio consisting of many fewer stocks must be created. One method to create the tracking portfolios is to find the best linear predictor of the index using a set of individual shares.

Let \mathbf{x}_i be the returns on a set of assets and let Y_i be the return on the index. The tracking error problem is to minimize the

$$E[(Y_i - \mathbf{X}_i \mathbf{w})^2]$$

where \mathbf{w} is a vector of portfolio weights. Portfolio tracking has the same structure as the best linear predictor and the optimal weights are $\hat{\mathbf{w}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

Data between January 5, 2010, and December 31, 2019, was used, a total of 2,515 trading days. The regression specification is simple: the return on the S&P is regressed on the returns on the sector

ETF returns,

$$R_i^{SP500} = \sum_{j=1}^{30} w_j R_{ij} + \varepsilon_i$$

where the portfolios are ordered alphabetically (not that this matters). The portfolio weights (which need not sum to 1) are presented in figure 3.16. All funds have positive weights, and the maximum just under 25%. More importantly, this portfolio has a correlation of 99.5% with the return on the S&P 500. Its return tracks the return of the S&P to within 1.4% per year. The tracking error variance is much smaller than the 14.7% annualized volatility of the S&P over this period.

While the regression estimates provide the solution to the unconditional tracking error problem, this estimator ignores two important considerations: how should stocks be selected, and how conditioning information (such as time-varying covariance) can be used. The first issue, which stocks to choose, is difficult and is typically motivated by the cost of trading and liquidity. The second issue will be re-examined using Multivariate GARCH and related models in a later chapter.

3.A Selected Proofs

Theorem 3.1.

$$\begin{aligned} E[\hat{\beta}|\mathbf{X}] &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}|\mathbf{X}] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon|\mathbf{X}] \\ &= \beta + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon|\mathbf{X}] \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\varepsilon|\mathbf{X}] \\ &= \beta \end{aligned}$$

□

Theorem 3.2.

$$\begin{aligned} V[\hat{\beta}|\mathbf{X}] &= E\left[\left(\hat{\beta} - E[\hat{\beta}|\mathbf{X}]\right)\left(\hat{\beta} - E[\hat{\beta}|\mathbf{X}]\right)'|\mathbf{X}\right] \\ &= E\left[\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)'|\mathbf{X}\right] \\ &= E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon\varepsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}|\mathbf{X}\right] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\varepsilon\varepsilon'|\mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{I}_n\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

□

Theorem 3.3. Without loss of generality $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X} + \mathbf{D}'$ where \mathbf{D}' must satisfy $\mathbf{D}'\mathbf{X} = \mathbf{0}$ and $E[\mathbf{D}'\boldsymbol{\varepsilon}|\mathbf{X}] = 0$ since

$$\begin{aligned} E[\tilde{\beta}|\mathbf{X}] &= E[\mathbf{C}\mathbf{y}|\mathbf{X}] \\ &= E\left[\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{D}'\right)(\mathbf{X}\beta + \boldsymbol{\varepsilon})|\mathbf{X}\right] \\ &= \beta + \mathbf{D}'\mathbf{X}\beta + E[\mathbf{D}'\boldsymbol{\varepsilon}|\mathbf{X}] \end{aligned}$$

and by assumption $\mathbf{C}\mathbf{y}$ is unbiased and so $E[\mathbf{C}\mathbf{y}|\mathbf{X}] = \beta$.

$$\begin{aligned} V[\tilde{\beta}|\mathbf{X}] &= E\left[\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{D}'\right)\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\left(\mathbf{D} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right)|\mathbf{X}\right] \\ &= E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}|\mathbf{X}\right] + E[\mathbf{D}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{D}|\mathbf{X}] + E\left[\mathbf{D}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}|\mathbf{X}\right] + E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{D}|\mathbf{X}\right] \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \sigma^2\mathbf{D}'\mathbf{D} + \sigma^2\mathbf{D}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}|\mathbf{X} + \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D} \\ &= V[\hat{\beta}|\mathbf{X}] + \sigma^2\mathbf{D}'\mathbf{D} + \mathbf{0} + \mathbf{0} \\ &= V[\hat{\beta}|\mathbf{X}] + \sigma^2\mathbf{D}'\mathbf{D} \end{aligned}$$

and so the variance of $\tilde{\beta}$ is equal to the variance of $\hat{\beta}$ plus a positive semi-definite matrix, and so

$$V[\tilde{\beta}|\mathbf{X}] - V[\hat{\beta}|\mathbf{X}] = \sigma^2\mathbf{D}'\mathbf{D} \geq \mathbf{0}$$

where the inequality is strict whenever $\mathbf{D} \neq \mathbf{0}$. □

Theorem 3.4.

$$\hat{\beta} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$$

and so $\hat{\beta}$ is a linear function of normal random variables $\boldsymbol{\varepsilon}$, and so it must be normal. Applying the results of Theorems 3.1 and 3.2 completes the proof. □

Theorem 3.5. □

$\hat{\beta} - \beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$ and $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{M}_X\mathbf{y} = \mathbf{M}_X\boldsymbol{\varepsilon}$, and so

$$\begin{aligned} E\left[\left(\hat{\beta} - \beta\right)\hat{\boldsymbol{\varepsilon}}'|\mathbf{X}\right] &= E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{M}_X|\mathbf{X}\right] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}]\mathbf{M}_X \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_X \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{M}_X\mathbf{X})' \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{0} \\ &= \mathbf{0} \end{aligned}$$

since $\mathbf{M}_X\mathbf{X} = \mathbf{0}$ by construction. $\hat{\beta}$ and $\hat{\boldsymbol{\varepsilon}}$ are jointly normally distributed since both are linear functions of $\boldsymbol{\varepsilon}$, and since they are uncorrelated they are independent.³¹

³¹Zero correlation is, in general, insufficient to establish that two random variables are independent. However, when two random variables are jointly normally distribution, they are independent if and only if they are uncorrelated.

Theorem 3.6. $\hat{\sigma}^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k}$ and so $(n-k)\hat{\sigma}^2 = \hat{\varepsilon}'\hat{\varepsilon}$. $\hat{\varepsilon} = \mathbf{M}_X\varepsilon$, so $(n-k)\hat{\sigma}^2 = \varepsilon'\mathbf{M}_X'\mathbf{M}_X\varepsilon$ and $(n-k)\frac{\hat{\sigma}^2}{\sigma^2} = \frac{\varepsilon'\mathbf{M}_X\varepsilon}{\sigma^2} = \frac{\varepsilon'\mathbf{M}_X\varepsilon}{\sigma^2} = \mathbf{z}'\mathbf{M}_X\mathbf{z}$ since \mathbf{M}_X is idempotent (and hence symmetric) where \mathbf{z} is a n by 1 multivariate normal vector with covariance \mathbf{I}_n . Finally, applying the result in Lemma 3.1, $\mathbf{z}'\mathbf{M}_X\mathbf{z} \sim \sum_{i=1}^n \lambda_i \chi_{1,i}^2$ where $\{\lambda_i\}$, $i = 1, 2, \dots, n$ are the eigenvalues of \mathbf{M}_X and $\chi_{1,i}^2$, $i = 1, 2, \dots, n$ are independent χ_1^2 random variables. Finally, note that \mathbf{M}_X is a rank $n-k$ idempotent matrix, so it must have $n-k$ eigenvalues equal to 1, $\lambda_i = 1$ for $i = 1, 2, \dots, n-k$ and k eigenvalues equal to 0, $\lambda_i = 0$ for $i = n-k+1, \dots, n$, and so the distribution is a χ_{n-k}^2 . \square

Lemma 3.1 (Quadratic Forms of Multivariate Normals). *Suppose $\mathbf{z} \sim N(\mathbf{0}, \Sigma)$ where Σ is a n by n positive semi-definite matrix, and let \mathbf{W} be a n by n positive semi-definite matrix, then*

$$\mathbf{z}'\mathbf{W}\mathbf{z} \sim N_2(\mathbf{0}, \Sigma; \mathbf{W}) \equiv \sum_{i=1}^n \lambda_i \chi_{1,i}^2$$

where λ_i are the eigenvalues of $\Sigma^{1/2}\mathbf{W}\Sigma^{1/2}$ and $N_2(\cdot)$ is known as a type-2 normal..

This lemma is a special case of Baldessari (1967) as presented in White (Lemma 8.2, 1996).

Theorem 3.8. The OLS estimator is the BUE estimator since it is unbiased by Theorem 3.1 and it achieves the Cramer-Rao lower bound (Theorem 3.7). \square

Theorem 3.9. Follows directly from the definition of a Student's t by applying Theorems 3.4, 3.5, and 3.2. \square

Theorem 3.10. Follows directly from the definition of a F_{V_1, V_2} by applying Theorems 3.4, 3.5, and 3.2. \square

Theorem 3.12.

$$\begin{aligned} \hat{\beta}_n - \beta &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\varepsilon \\ &= \left(\sum_{i=1}^n \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \sum_{i=1}^n \mathbf{x}_i' \varepsilon_i \\ &= \left(\frac{\sum_{i=1}^n \mathbf{x}_i' \mathbf{x}_i}{n} \right)^{-1} \frac{\sum_{i=1}^n \mathbf{x}_i' \varepsilon_i}{n} \end{aligned}$$

Since $E[\mathbf{x}_i' \mathbf{x}_i]$ is positive definite by Assumption 3.8, and $\{\mathbf{x}_i\}$ is stationary and ergodic by Assumption 3.7, then $\frac{\sum_{i=1}^n \mathbf{x}_i' \mathbf{x}_i}{n}$ will be positive definite for n sufficiently large, and so $\hat{\beta}_n$ exists. Applying the Ergodic Theorem (Theorem 3.21), $\frac{\sum_{i=1}^n \mathbf{x}_i' \mathbf{x}_i}{n} \xrightarrow{a.s.} \Sigma_{XX}$ and $\frac{\sum_{i=1}^n \mathbf{x}_i' \varepsilon_i}{n} \xrightarrow{a.s.} \mathbf{0}$ and by the Continuous Mapping Theorem (Theorem 3.22) combined with the continuity of the matrix inverse function, $\left(\frac{\sum_{i=1}^n \mathbf{x}_i' \mathbf{x}_i}{n} \right)^{-1} \xrightarrow{a.s.} \Sigma_{XX}^{-1}$, and so

$$\begin{aligned} \hat{\beta}_n - \beta &= \left(\frac{\sum_{i=1}^n \mathbf{x}_i' \mathbf{x}_i}{n} \right)^{-1} \frac{\sum_{i=1}^n \mathbf{x}_i' \varepsilon_i}{n} \\ &\xrightarrow{a.s.} \Sigma_{XX}^{-1} \cdot \mathbf{0} \\ &\xrightarrow{a.s.} \mathbf{0}. \end{aligned}$$

Finally, almost sure convergence implies convergence in probability and so $\hat{\beta}_n - \beta \xrightarrow{p} 0$ or $\hat{\beta}_n \xrightarrow{p} \beta$. \square

Theorem 3.21 (Ergodic Theorem). *If $\{\mathbf{z}_t\}$ is ergodic and its r^{th} moment, μ_r , is finite, then*

$$T^{-1} \sum_{t=1}^T \mathbf{z}_t^r \xrightarrow{a.s.} \mu_r$$

Theorem 3.22 (Continuous Mapping Theorem). *Given $\mathbf{g} : \mathbb{R}^k \rightarrow \mathbb{R}^l$, and any sequence of random k by l vectors $\{\mathbf{z}_n\}$ such that $\mathbf{z}_n \xrightarrow{a.s.} \mathbf{z}$ where \mathbf{z} is k by l , if \mathbf{g} is continuous at \mathbf{z} , then $\mathbf{g}(\mathbf{z}_n) \xrightarrow{a.s.} \mathbf{g}(\mathbf{z})$.*

Theorem 3.13. See White (Theorem 5.25, 2000). \square

Theorem 3.15.

$$\begin{aligned} \frac{\hat{\varepsilon}' \hat{\varepsilon}}{n} &= \frac{(\mathbf{y} - \mathbf{X}\hat{\beta}_n)' (\mathbf{y} - \mathbf{X}\hat{\beta}_n)}{n} \\ &= \frac{(\mathbf{y} - \mathbf{X}\hat{\beta}_n)' (\mathbf{y} - \mathbf{X}\hat{\beta}_n)}{n} \\ &= \frac{(\mathbf{y} - \mathbf{X}\hat{\beta}_n + \mathbf{X}\beta - \mathbf{X}\beta)' (\mathbf{y} - \mathbf{X}\hat{\beta}_n + \mathbf{X}\beta - \mathbf{X}\beta)}{n} \\ &= \frac{(\mathbf{y} - \mathbf{X}\beta + \mathbf{X}(\beta - \hat{\beta}_n))' (\mathbf{y} - \mathbf{X}\beta + \mathbf{X}(\beta - \hat{\beta}_n))}{n} \\ &= \frac{(\varepsilon + \mathbf{X}(\beta - \hat{\beta}_n))' (\varepsilon + \mathbf{X}(\beta - \hat{\beta}_n))}{n} \\ &= \frac{\varepsilon' \varepsilon}{n} + 2 \frac{(\beta - \hat{\beta}_n)' \mathbf{X}' \varepsilon}{n} + \frac{(\beta - \hat{\beta}_n)' \mathbf{X}' \mathbf{X} (\beta - \hat{\beta}_n)}{n} \end{aligned}$$

By the Ergodic Theorem and the existence of $E[\varepsilon_i^2]$ (Assumption 3.10), the first term converged to σ^2 . The second term

$$\frac{(\beta - \hat{\beta}_n)' \mathbf{X}' \varepsilon}{n} = (\beta - \hat{\beta}_n)' \frac{\sum_{i=1}^n \mathbf{X}' \varepsilon}{n} \xrightarrow{p} \mathbf{0}' \mathbf{0} = 0$$

since $\hat{\beta}_n$ is consistent and $E[\mathbf{x}_i \varepsilon_i] = 0$ combined with the Ergodic Theorem. The final term

$$\begin{aligned} \frac{(\beta - \hat{\beta}_n)' \mathbf{X}' \mathbf{X} (\beta - \hat{\beta}_n)}{n} &= (\beta - \hat{\beta}_n)' \frac{\mathbf{X}' \mathbf{X}}{n} (\beta - \hat{\beta}_n) \\ &\xrightarrow{p} \mathbf{0}' \Sigma_{\mathbf{XX}} \mathbf{0} = 0 \end{aligned}$$

and so the variance estimator is consistent. \square

Theorem 3.17.

$$\begin{aligned}\hat{\beta}_{1n} &= \left(\frac{\mathbf{X}'_1 \mathbf{X}_1}{n}\right)^{-1} \frac{\mathbf{X}'_1 \mathbf{y}}{n} \\ \left(\frac{\mathbf{X}'_1 \mathbf{X}_1}{n}\right)^{-1} \frac{\mathbf{X}'_1 (\mathbf{X}_1 + \mathbf{X}_2 + \varepsilon)}{n} &= \left(\frac{\mathbf{X}'_1 \mathbf{X}_1}{n}\right)^{-1} \frac{\mathbf{X}'_1 \mathbf{X}_1}{n} + \left(\frac{\mathbf{X}'_1 \mathbf{X}_1}{n}\right)^{-1} \frac{\mathbf{X}'_1 \mathbf{X}_2}{n} + \left(\frac{\mathbf{X}'_1 \mathbf{X}_1}{n}\right)^{-1} \frac{\mathbf{X}'_1 \varepsilon}{n} \\ &\xrightarrow{p} \beta_1 + \Sigma_{\mathbf{x}_1 \mathbf{x}_1}^{-1} \Sigma_{\mathbf{x}_1 \mathbf{x}_2} \beta_2 + \Sigma_{\mathbf{x}_1 \mathbf{x}_1}^{-1} \mathbf{0} \\ &= \beta_1 + \Sigma_{\mathbf{x}_1 \mathbf{x}_1}^{-1} \Sigma_{\mathbf{x}_1 \mathbf{x}_2} \beta_2\end{aligned}$$

where $\left(\frac{\mathbf{X}'_1 \mathbf{X}_1}{n}\right)^{-1} \xrightarrow{p} \Sigma_{\mathbf{x}_1 \mathbf{x}_1}^{-1}$ and $\frac{\mathbf{X}'_1 \mathbf{X}_1}{n} \xrightarrow{p} \Sigma_{\mathbf{x}_1 \mathbf{x}_2}$ by the Ergodic and Continuous Mapping Theorems (Theorems 3.21 and 3.22). Finally note that

$$\begin{aligned}\left(\frac{\mathbf{X}'_1 \mathbf{X}_1}{n}\right)^{-1} \frac{\mathbf{X}'_1 \mathbf{X}_2}{n} &= \left(\frac{\mathbf{X}'_1 \mathbf{X}_1}{n}\right)^{-1} [\mathbf{X}_1 \mathbf{x}_{2,1} \mathbf{X}_1 \mathbf{x}_{2,2} \dots \mathbf{X}_1 \mathbf{x}_{2,k_2}] \\ &= \left[\left(\frac{\mathbf{X}'_1 \mathbf{X}_1}{n}\right)^{-1} \mathbf{X}_1 \mathbf{x}_{2,1} \left(\frac{\mathbf{X}'_1 \mathbf{X}_1}{n}\right)^{-1} \mathbf{X}_1 \mathbf{x}_{2,2} \dots \left(\frac{\mathbf{X}'_1 \mathbf{X}_1}{n}\right)^{-1} \mathbf{X}_1 \mathbf{x}_{2,k_2} \right] \\ &= [\hat{\delta}_{1n} \hat{\delta}_{2n} \dots \hat{\delta}_{k_2 n}]\end{aligned}$$

where δ_j is the regression coefficient in $\mathbf{x}_{2,j} = \mathbf{X} \delta_j + \eta_j$. □

Theorem 3.18. See White (Theorem 6.3, 2000). □

Theorem 3.19. See White (Theorem 6.4, 2000). □

Theorem 3.20. By Assumption 3.15,

$$\mathbf{V}^{-\frac{1}{2}} \mathbf{y} = \mathbf{V}^{-\frac{1}{2}} \mathbf{X} \beta + \mathbf{V}^{-\frac{1}{2}} \varepsilon$$

and $\mathbf{V} \left[\mathbf{V}^{-\frac{1}{2}} \varepsilon \right] = \sigma^2 \mathbf{I}_n$, uncorrelated and homoskedastic, and so Theorem 3.3 can be applied. □

Shorter Problems

Problem 3.1. Derive the OLS estimator for the model $Y_i = \alpha + \varepsilon_i$.

Problem 3.2. Derive the OLS estimator for the model $Y_i = \beta X_i + \varepsilon_i$.

Problem 3.3. What are information criteria and how are they used?

Problem 3.4. Outline the steps to compute the bootstrap variance estimator for a regression when the data are heteroskedastic.

Problem 3.5. Discuss White's covariance estimator, and in particular when should White's covariance estimator be used? What are the consequences to using White's covariance estimator when it is not needed? How can one determine if White's covariance estimator is needed?

Problem 3.6. Suppose $Z_i = a + bX_i$, and two models are estimated using OLS: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ and $Y_i = \gamma_0 + \gamma_1 Z_i + \eta_i$, What the relationship between γ and β and between $\hat{\varepsilon}_i$ and $\hat{\eta}_i$?

Problem 3.7. Describe the steps to implement k -fold cross-validation in a regression to select a model.

Longer Exercises

Exercise 3.1. Imagine you have been given the task of evaluating the relationship between the return on a mutual fund and the number of years its manager has been a professional. You have a panel data set which covers all of the mutual funds returns in the year 1970-2005. Consider the regression

$$R_{i,t} = \alpha + \beta \text{exper}_{i,t} + \varepsilon_{i,t}$$

where r_{it} is the return on fund i in year t and exper_{it} is the number of years the fund manager has held her job in year t . The initial estimates of β and α are computed by stacking all of the observations into a vector and running a single OLS regression (across all funds and all time periods).

1. What test statistic would you use to determine whether experience has a positive effect?
2. What are the null and alternative hypotheses for the above test?
3. What does it mean to make a type I error in the above test? What does it mean to make a type II error in the above test?
4. Suppose that experience has no effect on returns but that unlucky managers get fired and thus do not gain experience. Is this a problem for the above test? If so, can you comment on its likely effect?
5. Could the estimated $\hat{\beta}$ ever be valid if mutual funds had different risk exposures? If so, why? If not, why not?
6. If mutual funds do have different risk exposures, could you write down a model which may be better suited to testing the effect of managerial experience than the initial simple specification? If it makes your life easier, you can assume there are only 2 mutual funds and 1 risk factor to control for.

Exercise 3.2. Consider the linear regression

$$Y_t = \beta X_t + \varepsilon_t$$

1. Derive the least-squares estimator. What assumptions are you making in the derivation of the estimator?
2. Under the classical assumptions, derive the variance of the estimator $\hat{\beta}$.
3. Suppose the errors ε_t have an AR(1) structure where $\varepsilon_t = \rho \varepsilon_{t-1} + \eta_t$ where $\eta_t \xrightarrow{d} N(0, 1)$ and $|\rho| < 1$. What is the variance of $\hat{\beta}$ now?
4. Now suppose that the errors have the same AR(1) structure but the x_t variables are i.i.d.. What is the variance of $\hat{\beta}$ now?
5. Finally, suppose the linear regression is now

$$Y_t = \alpha + \beta X_t + \varepsilon_t$$

where ε_t has an AR(1) structure and that x_t is i.i.d.. What is the covariance of $[\alpha \ \beta]'$?

Exercise 3.3. Consider the simple regression model $Y_i = \beta X_{1,i} + \varepsilon_i$ where the random error terms are i.i.d. with mean zero and variance σ^2 and are uncorrelated with the $X_{1,i}$.

1. Show that the OLS estimator of β is consistent.
2. Is the previously derived OLS estimator of β still consistent if $Y_i = \alpha + \beta X_{1,i} + \varepsilon_i$? Show why or why not.
3. Now suppose the data generating process is

$$Y_i = \beta_1 X_{1,i} + \beta_2 X_{2,i} + \varepsilon_i$$

Derive the OLS estimators of β_1 and β_2 .

4. Derive the asymptotic covariance of this estimator using the method of moments approach.
 - (a) What are the moment conditions?
 - (b) What is the Jacobian?
 - (c) What does the Jacobian limit to? What does this require?
 - (d) What is the covariance of the moment conditions. Be as general as possible.

In all of the above, clearly state any additional assumptions needed.

Exercise 3.4. Let $\hat{\mathbf{S}}$ be the sample covariance matrix of $\mathbf{z} = [\mathbf{y} \ \mathbf{X}]$, where \mathbf{X} does not include a constant

$$\hat{\mathbf{S}} = n^{-1} \sum_{i=1}^n (\mathbf{z}_i - \bar{\mathbf{z}})'(\mathbf{z}_i - \bar{\mathbf{z}})$$

$$\hat{\mathbf{S}} = \begin{bmatrix} \hat{s}_{yy} & \hat{s}'_{xy} \\ \hat{s}_{xy} & \hat{\mathbf{S}}_{xx} \end{bmatrix}$$

and suppose n , the sample size, is known ($\hat{\mathbf{S}}$ is the sample covariance estimator). Under the small-sample assumptions (including homoskedasticity and normality if needed), describe one method, using only $\hat{\mathbf{S}}$, $\bar{\mathbf{X}}$ (the 1 by $k-1$ sample mean of the matrix \mathbf{X} , column-by-column), \bar{y} and n , to

1. Estimate $\hat{\beta}_1, \dots, \hat{\beta}_k$ from a model

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \dots + \beta_k X_{k,i} + \varepsilon_i$$

2. Estimate s , the standard error of the regression

3. Test $H_0 : \beta_j = 0, j = 2, \dots, k$

Exercise 3.5. Consider the regression model

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$$

where the random error terms are i.i.d. with mean zero and variance σ^2 and are uncorrelated with the x_i . Also assume that x_i is i.i.d. with mean μ_x and variance σ_x^2 , both finite.

1. Using scalar notation, derive the OLS estimators of β_1 and β_2 .
2. Show these estimators are consistent. Are any further assumptions needed?
3. Show that the matrix expression for the estimator of the regression parameters, $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, is identical to the estimators derived using scalar notation.

Exercise 3.6. Let $\mathbf{x}_m\beta$ be the best linear projection of Y_m . Let ε_m be the prediction error.

1. What is the variance of a projected Y ?
2. What is the variance if the β s are estimated using regressors that do not include observation m (and hence not \mathbf{x}_m or ε_m)? Hint: You can use any assumptions in the notes, just be clear what you are assuming.

Exercise 3.7. Are Wald tests of linear restrictions in a linear regression invariant to linear reparameterizations? Hint: Let \mathbf{F} be an invertible matrix. Parameterize W in the case where $H_0 : \mathbf{R}\beta - \mathbf{r} = \mathbf{0}$ and $H_0 : \mathbf{F}(\mathbf{R}\beta - \mathbf{r}) = \mathbf{F}\mathbf{R}\beta - \mathbf{F}\mathbf{r} = \mathbf{0}$.

1. Are they the same?
2. Show that $n \cdot \mathbf{R}^2$ has an asymptotic χ_{k-1}^2 distribution under the classical assumptions when the model estimated is

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \dots + \beta_k X_{k,i} + \varepsilon_i$$

Hint: What is the does the distribution of c/v converge to as $v \rightarrow \infty$ when $c \sim \chi_v^2$.

Exercise 3.8. Suppose an unrestricted model is

$$Y_i = \beta_1 + \beta_2 X_{1,i} + \beta_3 X_{2,i} + \beta_4 X_{3,i} + \varepsilon_i$$

1. Sketch the steps required to test a null $H_0 : \beta_2 = \beta_3 = 0$ in the large-sample framework using a Wald test and an LM test.
2. Sketch the steps required to test a null $H_0 : \beta_2 + \beta_3 + \beta_4 = 1$ in the small-sample framework using a Wald test, a t -test, an LR test, and an LM test.

In the above questions be clear what the null and alternative are, which regressions must be estimated, how to compute any numbers that are needed and the distribution of the test statistic.

Exercise 3.9. Let Y_i and X_i conform to the small-sample assumptions and let $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$. Define another estimator

$$\check{\beta}_2 = \frac{\bar{Y}_H - \bar{Y}_L}{\bar{X}_H - \bar{X}_L}$$

where \bar{X}_H is the average value of X_i given $X_i > \text{median}(\mathbf{x})$, and \bar{Y}_H is the average value of Y_i for n such that $X_i > \text{median}(\mathbf{x})$. \bar{X}_L is the average value of X_i given $X_i \leq \text{median}(\mathbf{x})$, and \bar{Y}_L is the average value of Y_i for n such that $X_i \leq \text{median}(\mathbf{x})$ (both \bar{X} and \bar{Y} depend on the order of X_i , and not Y_i). For example, suppose the X_i were ordered such that $X_1 < X_2 < X_3 < \dots < X_i$ and n is even. Then,

$$\bar{X}_L = \frac{2}{n} \sum_{i=1}^{n/2} X_i$$

and

$$\bar{X}_H = \frac{2}{n} \sum_{i=n/2+1}^n X_i$$

1. Is $\check{\beta}_2$ unbiased, conditional on \mathbf{X} ?
2. Is $\check{\beta}_2$ consistent? Are any additional assumptions needed beyond those of the small-sample framework?
3. What is the variance of $\check{\beta}_2$, conditional on \mathbf{X} ?

Exercise 3.10. Suppose

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

and that variable Z_i is available where $V[Z_i] = \sigma_z^2 > 0$, $\text{Corr}(X_i, Z_i) = \rho \neq 0$ and $E[\varepsilon_i | \mathbf{z}] = 0$, $n = 1, \dots, N$. Further suppose the other assumptions of the small-sample framework hold. Rather than the usual OLS estimator,

$$\check{\beta}_2 = \frac{\sum_{i=1}^n (Z_i - \bar{Z}) Y_i}{\sum_{i=1}^n (Z_i - \bar{Z}) X_i}$$

is used.

1. Is $\check{\beta}_2$ a reasonable estimator for β_2 ?
2. What is the variance of $\check{\beta}_2$, conditional on \mathbf{x} and \mathbf{z} ?
3. What does the variance limit to (i.e., not conditioning on \mathbf{x} and \mathbf{z})?
4. How is this estimator related to OLS, and what happens to its variance when OLS is used (Hint: What is $\text{Corr}(X_i, X_i)$?)

Exercise 3.11. Let $\{Y_i\}_{i=1}^n$ and $\{X_i\}_{i=1}^n$ conform to the small-sample assumptions and let $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$. Define the estimator

$$\check{\beta}_2 = \frac{\bar{Y}_H - \bar{Y}_L}{\bar{X}_H - \bar{X}_L}$$

where \bar{X}_H is the average value of X_i given $X_i > \text{median}(\mathbf{x})$, and \bar{Y}_H is the average value of Y_i for i such that $X_i > \text{median}(\mathbf{x})$. \bar{X}_L is the average value of X_i given $X_i \leq \text{median}(\mathbf{x})$, and \bar{Y}_L is the average value of Y_i for i such that $X_i \leq \text{median}(\mathbf{x})$ (both \bar{X} and \bar{Y} depend on the order of X_i , and not Y_i). For example, suppose the X_i were ordered such that $X_1 < X_2 < X_3 < \dots < X_n$ and n is even. Then,

$$\bar{X}_L = \frac{2}{n} \sum_{i=1}^{n/2} X_i$$

and

$$\bar{X}_H = \frac{2}{n} \sum_{i=n/2+1}^n X_i$$

1. Is $\check{\beta}_2$ unbiased, conditional on \mathbf{X} ?

2. Is $\check{\beta}_2$ consistent? Are any additional assumptions needed beyond those of the small-sample framework?
3. What is the variance of $\check{\beta}_2$, conditional on \mathbf{X} ?

Next consider the estimator

$$\ddot{\beta}_2 = \frac{\bar{Y}}{\bar{X}}$$

where \bar{Y} and \bar{X} are sample averages of $\{Y_i\}$ and $\{X_i\}$, respectively.

4. Is $\ddot{\beta}_2$ unbiased, conditional on \mathbf{X} ?
5. Is $\ddot{\beta}_2$ consistent? Are any additional assumptions needed beyond those of the small-sample framework?
6. What is the variance of $\ddot{\beta}_2$, conditional on \mathbf{X} ?

Exercise 3.12. Suppose an unrestricted model is

$$Y_i = \beta_1 + \beta_2 X_{1,i} + \beta_3 X_{2,i} + \beta_4 X_{3,i} + \varepsilon_i$$

1. Discuss which features of estimators each of the three major tests, Wald, Likelihood Ratio, and Lagrange Multiplier, utilize in testing.
2. Sketch the steps required to test a null $H_0 : \beta_2 = \beta_3 = 0$ in the large-sample framework using Wald, LM, and LR tests.
3. What are type I & II errors?
4. What is the size of a test?
5. What is the power of a test?
6. What influences the power of a test?
7. What is the most you can say about the relative power of a Wald, LM, and LR test of the same null?

Exercise 3.13. Consider the regression model

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$$

where the random error terms are i.i.d. with mean zero and variance σ^2 and are uncorrelated with the X_i . Also assume that X_i is i.i.d. with mean μ_x and variance σ_x^2 , both finite.

1. Using scalar notation, derive the OLS estimators of β_1 and β_2 .
2. Why are these estimators are consistent? Are any further assumptions needed?
3. Show that the matrix expression for the estimator of the regression parameters, $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, is identical to the estimators derived using scalar notation.

4. Suppose instead

$$Y_i = \gamma_1 + \gamma_2 (X_i - \bar{X}) + \varepsilon_i$$

was fit to the data. How are the estimates of the γ s related to the β s?

5. What can you say about the relationship between the t -statistics of the γ s and the β s?
6. How would you test for heteroskedasticity in the regression?
7. Since the errors are i.i.d. there is no need to use White's covariance estimator for this regression. What are the consequences of using White's covariance estimator if it is not needed?

Exercise 3.14. Suppose $Y_i = \alpha + \beta X_i + \varepsilon_i$ where $E[\varepsilon_i | X] = 0$ and $V[\varepsilon_i] = \sigma^2$ for all i .

1. Derive the OLS estimators of α and β .
2. Describe the trade-offs when deciding whether to use the classic parameter covariance estimator, $\hat{\sigma}^2 \Sigma_{XX}^{-1}$, and White's parameter covariance estimator, $\Sigma_{XX}^{-1} \mathbf{S} \Sigma_{XX}^{-1}$?
3. Describe a procedure to formally test whether White's covariance estimator is required.
4. Suppose the true model is as above, but instead the model $Y_i = \gamma + \varepsilon_i$ is fit. What is the most you can say about the the OLS estimate of $\hat{\gamma}$?
5. What is Winsorization in the context of a regression, and how is it useful?

Exercise 3.15. Consider the APT regression

$$R_t^e = \alpha + \beta_m R_{m,t}^e + \beta_s R_{smb,t} + \beta_v R_{hml,t} + \varepsilon_t$$

where $R_{m,t}^e$ is the excess return on the market, $R_{smb,t}$ is the return on the size factor, $R_{hml,t}$ is the return on value factor and R_t^e is an excess return on a portfolio of assets. Using the information provided in the tables below below, answer the following questions:

1. Is there evidence that this portfolio is market neutral?
2. Are the size and value factors needed in this portfolio to adequately capture the cross-sectional dynamics?
3. Is there evidence of conditional heteroskedasticity in this model?
4. What are the trade-offs for choosing a covariance estimator for making inference on this model?
5. Define the size and power of a statistical test.
6. What factors affect the power of a statistical test?
7. Outline the steps to implement the correct bootstrap covariance estimator for these parameters. Justify the method you chose using the information provided.

Notes: All models were estimated on $n = 100$ data points. Models 1 and 2 correspond to the specification above. In model 1 R_{smb} and R_{hml} have been excluded. Model 3, 4 and 5 are all version of

$$\begin{aligned}\hat{\varepsilon}_t^2 = & \gamma_0 + \gamma_1 R_{m,t}^e + \gamma_2 R_{smb,t} + \gamma_3 R_{hml,t} + \gamma_4 (R_{m,t}^e)^2 + \gamma_5 R_{m,t}^e R_{smb,t} \\ & + \gamma_6 R_{m,t}^e R_{hml,t} + \gamma_7 R_{smb,t}^2 + \gamma_8 R_{smb,t} R_{hml,t} + \gamma_9 R_{hml,t}^2 + \eta_t\end{aligned}$$

$\hat{\varepsilon}_t$ was computed using Model 1 for the results under Model 3, and using model 2 for the results under Models 4 and 5. R^2 is the R-squared and n is the number of observations.

Parameter Estimates

	Model 1	Model 2	Model 3	Model 4	Model 5
α	0.128	0.089	γ_0	0.984	0.957
β_m	1.123	0.852	γ_1	-0.779	-0.498
β_{smb}		0.600	γ_2		-0.046
β_{hml}		-0.224	γ_3		0.124
			γ_4	0.497	0.042
			γ_5		0.295
			γ_6		0.049
			γ_7		0.684
			γ_8		0.036
			γ_9		-0.149
					-0.362
					-0.005
					0.128
R^2	0.406	0.527		0.134	0.126
					0.037

Parameter Covariance Estimates

The estimated covariance matrices from the asymptotic distribution

$$\sqrt{n} (\hat{\beta} - \hat{\beta}_0) \xrightarrow{d} N(0, C)$$

are below where C is either $\hat{\sigma}^2 \hat{\Sigma}_{XX}^{-1}$ or $\hat{\Sigma}_{XX}^{-1} \hat{S} \hat{\Sigma}_{XX}^{-1}$.

CAP-M

$$\hat{\sigma}^2 \hat{\Sigma}_{XX}^{-1}$$

	α	β_m
α	1.365475	0.030483
β_m	0.030483	1.843262

$$\hat{\Sigma}_{XX}^{-1} \hat{S} \hat{\Sigma}_{XX}^{-1}$$

	α	β_m
α	1.341225	-0.695235
β_m	-0.695235	2.747142

Fama-French Model

$$\hat{\sigma}^2 \hat{\Sigma}_{XX}^{-1}$$

	α	β_m	β_{smb}	β_{hml}
α	1.100680	0.103611	-0.088259	-0.063529
β_m	0.103611	1.982761	-0.619139	-0.341118
β_{smb}	-0.088259	-0.619139	1.417318	-0.578388
β_{hml}	-0.063529	-0.341118	-0.578388	1.686200

$$\hat{\Sigma}_{XX}^{-1} \hat{S} \hat{\Sigma}_{XX}^{-1}$$

	α	β_m	β_{smb}	β_{hml}
α	1.073227	-0.361618	-0.072784	0.045732
β_m	-0.361618	2.276080	-0.684809	0.187441
β_{smb}	-0.072784	-0.684809	1.544745	-1.074895
β_{hml}	0.045732	0.187441	-1.074895	1.947117

χ_m^2 critical values

Critical value for a 5% test when the test statistic has a χ_m^2 distribution.

m	1	2	3	4	8	9	10
Crit Val.	3.84	5.99	7.81	9.48	15.50	16.91	18.30
m	90	91	98	99	100		
Crit Val.	113.14	114.26	122.10	123.22	124.34		

Matrix Inverse

The inverse of a 2 by 2 matrix

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Exercise 3.16. Suppose $Y_i = \alpha + \beta X_i + \varepsilon_i$ where $E[\varepsilon_i|X] = 0$ and $V[\varepsilon_i] = \sigma^2$ for all i .

1. Describe the trade-offs when deciding whether to use the classic parameter covariance estimator, $\hat{\sigma}^2 \Sigma_{XX}^{-1}$, and White's parameter covariance estimator, $\Sigma_{XX}^{-1} \mathbf{S} \Sigma_{XX}^{-1}$?
2. Describe a procedure to formally test whether White's covariance estimator is required.
3. Suppose the true model is as above, but instead the model $Y_i = \gamma + \varepsilon_i$ is fit. What is the most you can say about the OLS estimate of $\hat{\gamma}$?
4. Define the size and power of a statistical test.
5. What factors affect the power of a statistical test?
6. What is Winsorization in the context of a regression, and how is it useful?

Chapter 4

Analysis of a Single Time Series

Note: The primary reference for these notes is Enders (2004). An alternative and more technical treatment can be found in Hamilton (1994).

Most data used in financial econometrics occur sequentially through time. Interest rates, asset returns, and foreign exchange rates are all examples of time series. This chapter introduces time-series econometrics and focuses primarily on linear models, although some common non-linear models are described in the final section. The analysis of time-series data begins by defining two key concepts in the analysis of time series: stationarity and ergodicity. The chapter next turns to Autoregressive Moving Average models (ARMA) and covers the structure of these models, stationarity conditions, model selection, estimation, inference, and forecasting. Finally, The chapter concludes by examining nonstationary time series.

4.1 Stochastic Processes

A stochastic process is an arbitrary sequence of random data and is denoted

$$\{y_t\} \tag{4.1}$$

where $\{\cdot\}$ is used to indicate that the y s form a sequence. The simplest non-trivial stochastic process specifies that $y_t \stackrel{\text{i.i.d.}}{\sim} D$ for some distribution D , for example normal. Another simple stochastic process is the random walk,

$$y_t = y_{t-1} + \varepsilon_t$$

where ε_t is an i.i.d. process.

4.2 Stationarity, Ergodicity, and the Information Set

Stationarity is a probabilistically meaningful measure of regularity. This regularity can be exploited to estimate unknown parameters and characterize the dependence between observations across time. If the data generating process frequently changed in an unpredictable manner, constructing a meaningful model would be difficult or impossible.

Stationarity exists in two forms, strict stationarity and covariance (also known as weak) stationarity. Covariance stationarity is important when modeling the mean of a process, although strict stationarity is useful in more complicated settings, such as non-linear models.

Definition 4.1 (Strict Stationarity). A stochastic process $\{y_t\}$ is strictly stationary if the joint distribution of $\{y_t, y_{t+1}, \dots, y_{t+h}\}$ only depends only on h and not on t .

Strict stationarity requires that the *joint* distribution of a stochastic process does not depend on time and so the only factor affecting the relationship between two observations is the gap between them. Strict stationarity is weaker than i.i.d. since the process may be dependent but it is nonetheless a strong assumption and implausible for many time series, including both financial and macroeconomic data.

Covariance stationarity, on the other hand, only imposes restrictions on the first two moments of a stochastic process.

Definition 4.2 (Covariance Stationarity). A stochastic process $\{y_t\}$ is covariance stationary if

$$\begin{aligned} E[y_t] &= \mu \quad \text{for } t = 1, 2, \dots \\ V[y_t] &= \sigma^2 < \infty \quad \text{for } t = 1, 2, \dots \\ E[(y_t - \mu)(y_{t-s} - \mu)] &= \gamma_s \quad \text{for } t = 1, 2, \dots, s = 1, 2, \dots, t-1. \end{aligned} \tag{4.2}$$

Covariance stationarity requires that both the unconditional mean and unconditional variance are finite and do not change with time. Note that covariance stationarity only applies to *unconditional moments* and not conditional moments, and so a covariance process may have a varying conditional mean (i.e. be predictable).

These two types of stationarity are related although neither nests the other. If a process is strictly stationary *and* has finite second moments, then it is covariance stationary. If a process is covariance stationary and the joint distribution of the studentized residuals (demeaned and standardized by their standard deviation) does not depend on time, then the process is strictly stationary. However, one type can occur without the other, both can occur or neither may be applicable to a particular time series. For example, if a process has higher order moments which depend on time (e.g., time-varying kurtosis), it may be covariance stationary but not strictly stationary. Alternatively, a sequence of i.i.d. Student's t random variables with 2 degrees of freedom is strictly stationary but not covariance stationary since the variance of a t_2 is infinite.

$\gamma_s = E[(y_t - \mu)(y_{t-s} - \mu)]$ is the covariance of y_t with itself at a different point in time, known as the s^{th} autocovariance. γ_0 is the lag-0 autocovariance, the same quantity as the *long-run* variance of y_t (i.e. $\gamma_0 = V[y_t]$).¹

Definition 4.3 (Autocovariance). The autocovariance of a covariance stationary scalar process $\{y_t\}$ is defined

$$\gamma_s = E[(y_t - \mu)(y_{t-s} - \mu)] \tag{4.3}$$

where $\mu = E[y_t]$. Note that $\gamma_0 = E[(y_t - \mu)(y_t - \mu)] = V[y_t]$.

Ergodicity is another important concept in the analysis of time series and is one form of asymptotic independence.

¹The use of long-run variance is used to distinguish $V[y_t]$ from the innovation variance, $V[\varepsilon_t]$, also known as the short-run variance.

Definition 4.4 (Ergodicity). Let $\{y_t\}$ be a stationary sequence. $\{y_t\}$ is ergodic if for any two bounded functions $f : \mathbb{R}^k \rightarrow \mathbb{R}$ $g : \mathbb{R}^l \rightarrow \mathbb{R}$

$$\begin{aligned} & \lim_{j \rightarrow \infty} |\mathbb{E}[f(y_t, \dots, y_{t+k}) g(y_{t+j}, \dots, y_{t+l+j})]| \\ &= |\mathbb{E}[f(y_t, \dots, y_{t+k})]| |\mathbb{E}[g(y_{t+j}, \dots, y_{t+l+j})]| \end{aligned} \quad (4.4)$$

In essence, if an ergodic stochastic process is sampled at two points far apart in time, these samples will be independent. The ergodic theorem provides a practical application of ergodicity.

Theorem 4.1 (Ergodic Theorem). If $\{y_t\}$ is ergodic and its r^{th} moment μ_r is finite, then $T^{-1} \sum_{t=1}^T y_t^r \xrightarrow{P} \mu_r$.

The ergodic theorem states that averages will converge to their expectation provided the expectation exists. The intuition for this results follows from the definition of ergodicity since samples far apart in time are (effectively) independent, and so errors average across time.

Not all series are ergodic. Let $y_t = \eta + \varepsilon_t$ where $\eta \sim N(0, 1)$, $\varepsilon_t \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$ and η and ε_t are independent for any t . Note that η is drawn only once (not every t). Clearly, $\mathbb{E}[y_t] = 0$. However, $T^{-1} \sum_{t=1}^T y_t \xrightarrow{P} \eta \neq 0$, and so even though the average converges it does not converge to $\mathbb{E}[y_t]$ since the effect of the initial draw of η is present in every observation of $\{y_t\}$.

The third important building block of time-series models is white noise. White noise generalizes i.i.d. noise and allows for dependence in a series as long as three conditions are satisfied: the series is mean zero, uncorrelated and has finite second moments.

Definition 4.5 (White Noise). A process $\{\varepsilon_t\}$ is known as white noise if

$$\begin{aligned} \mathbb{E}[\varepsilon_t] &= 0 \quad \text{for } t = 1, 2, \dots \\ \mathbb{V}[\varepsilon_t] &= \sigma^2 < \infty \quad \text{for } t = 1, 2, \dots \\ \mathbb{E}[\varepsilon_t \varepsilon_{t-j}] &= \text{Cov}(\varepsilon_t, \varepsilon_{t-j}) = 0 \quad \text{for } t = 1, 2, \dots, j \neq 0. \end{aligned} \quad (4.5)$$

An i.i.d. series with finite second moments is trivially white noise, but other important processes, such as residuals following an ARCH (Autoregressive Conditional Heteroskedasticity) process, may also be white noise although not independent since white noise only requires linear independence.² A white noise process is also covariance stationary since it satisfies all three conditions: the mean, variance, and autocovariances are all finite and do not depend on time.

The final important concepts are conditional expectation and the information set. The information set at time t is denoted \mathcal{F}_t and contains all time t measurable events³, and so the information set includes realization of all variables which have occurred on or before t . For example, the information set for January 3, 2008 contains all stock returns up to an including those which occurred on January 3. It also includes everything else known at this time such as interest rates, foreign exchange rates or the scores of recent football games. Many expectations will often be made conditional on the time- t information set, expressed $\mathbb{E}[y_{t+h} | \mathcal{F}_t]$, or in abbreviated form as $E_t[y_{t+h}]$. The conditioning information set matters when taking expectations and $\mathbb{E}[y_{t+h}]$, $E_t[y_{t+h}]$ and $E_{t+h}[y_{t+h}]$ are not the same. Conditional variance is similarly defined, $\mathbb{V}[y_{t+h} | \mathcal{F}_t] = V_t[y_{t+h}] = E_t[(y_{t+h} - E_t[y_{t+h}])^2]$.

²Residuals generated from an ARCH process have dependence in conditional variances but not mean.

³A measurable event is any event that can have probability assigned to it at time t . In general this includes any observed variable but can also include time t beliefs about latent (unobserved) variables such as volatility or the final revision of the current quarter's GDP.

4.3 ARMA Models

Autoregressive moving average (ARMA) processes form the core of time-series analysis. The ARMA class can be decomposed into two smaller classes, autoregressive (AR) processes and moving average (MA) processes.

4.3.1 Moving Average Processes

The 1storder moving average, written MA(1), is the simplest non-degenerate time-series process,

$$y_t = \phi_0 + \theta_1 \varepsilon_{t-1} + \varepsilon_t$$

where ϕ_0 and θ_1 are parameters and ε_t a white noise series. This process stipulates that the current value of y_t depends on both a new shock and the previous shock. For example, if θ is negative, the current realization will “bounce back” from the previous shock.

Definition 4.6 (First Order Moving Average Process). A first order Moving Average process (MA(1)) has dynamics which follow

$$y_t = \phi_0 + \theta_1 \varepsilon_{t-1} + \varepsilon_t \quad (4.6)$$

where ε_t is a white noise process with the additional property that $E_{t-1}[\varepsilon_t] = 0$.

It is simple to derive both the conditional and unconditional means in this process. The conditional mean is

$$\begin{aligned} E_{t-1}[y_t] &= E_{t-1}[\phi_0 + \theta_1 \varepsilon_{t-1} + \varepsilon_t] \\ &= \phi_0 + \theta_1 E_{t-1}[\varepsilon_{t-1}] + E_{t-1}[\varepsilon_t] \\ &= \phi_0 + \theta_1 \varepsilon_{t-1} + 0 \\ &= \phi_0 + \theta_1 \varepsilon_{t-1} \end{aligned} \quad (4.7)$$

where $E_{t-1}[\varepsilon_t] = 0$ follows by assumption that the shock is unpredictable using the time- $t - 1$ information set, and since ε_{t-1} is in the time- $t - 1$ information set ($\varepsilon_{t-1} \in \mathcal{F}_{t-1}$), it passes through the time- $t - 1$ conditional expectation. The unconditional mean is

$$\begin{aligned} E[y_t] &= E[\phi_0 + \theta_1 \varepsilon_{t-1} + \varepsilon_t] \\ &= \phi_0 + \theta_1 E[\varepsilon_{t-1}] + E[\varepsilon_t] \\ &= \phi_0 + \theta_1 0 + 0 \\ &= \phi_0. \end{aligned} \quad (4.8)$$

Comparing these two results, the unconditional mean of y_t , $E[y_t]$, is ϕ_0 while the conditional mean $E_{t-1}[y_t] = \phi_0 + \theta_1 \varepsilon_{t-1}$. This difference reflects the persistence of the previous shock in the current period. The variances can be similarly derived,

$$\begin{aligned}
V[y_t] &= E[(\phi_0 + \theta_1 \varepsilon_{t-1} + \varepsilon_t - E[\phi_0 + \theta_1 \varepsilon_{t-1} + \varepsilon_t])^2] \\
&= E[(\phi_0 + \theta_1 \varepsilon_{t-1} + \varepsilon_t - \phi_0)^2] \\
&= E[(\theta_1 \varepsilon_{t-1} + \varepsilon_t)^2] \\
&= \theta_1^2 E[\varepsilon_{t-1}^2] + E[\varepsilon_t^2] + 2\theta_1 E[\varepsilon_{t-1} \varepsilon_t] \\
&= \sigma^2 \theta_1^2 + \sigma^2 + 0 \\
&= \sigma^2 (1 + \theta_1^2)
\end{aligned} \tag{4.9}$$

where $E[\varepsilon_{t-1} \varepsilon_t]$ follows from the white noise assumption. The conditional variance is

$$\begin{aligned}
V_{t-1}[y_t] &= E_{t-1}\left[(\phi_0 + \theta_1 \varepsilon_{t-1} + \varepsilon_t - E_{t-1}[\phi_0 + \theta_1 \varepsilon_{t-1} + \varepsilon_t])^2\right] \\
&= E_{t-1}\left[(\phi_0 + \theta_1 \varepsilon_{t-1} + \varepsilon_t - \phi_0 - \theta_1 \varepsilon_{t-1})^2\right] \\
&= E_{t-1}[\varepsilon_t^2] \\
&= \sigma_t^2
\end{aligned} \tag{4.10}$$

where σ_t^2 is the conditional variance of ε_t . White noise does not have to be homoskedastic, although if ε_t is homoskedastic then $V_{t-1}[y_t] = E[\sigma_t^2] = \sigma^2$. Like the mean, the unconditional variance and the conditional variance are different. The unconditional variance is unambiguously larger than the average conditional variance which reflects the extra variability introduced by the moving average term.

Finally, the autocovariance can be derived

$$\begin{aligned}
E[(y_t - E[y_t])(y_{t-1} - E[y_{t-1}])] &= E[(\phi_0 + \theta_1 \varepsilon_{t-1} + \varepsilon_t - \phi_0)(\phi_0 + \theta_1 \varepsilon_{t-2} + \varepsilon_{t-1} - \phi_0)] \\
&= E[\theta_1 \varepsilon_{t-1}^2 + \theta_1 \varepsilon_t \varepsilon_{t-2} + \varepsilon_t \varepsilon_{t-1} + \theta_1^2 \varepsilon_{t-1} \varepsilon_{t-2}] \\
&= \theta_1 E[\varepsilon_{t-1}^2] + \theta_1 E[\varepsilon_t \varepsilon_{t-2}] + E[\varepsilon_t \varepsilon_{t-1}] + \theta_1^2 E[\varepsilon_{t-1} \varepsilon_{t-2}] \\
&= \theta_1 \sigma^2 + 0 + 0 + 0 \\
&= \theta_1 \sigma^2
\end{aligned} \tag{4.11}$$

$$\begin{aligned}
E[(y_t - E[y_t])(y_{t-2} - E[y_{t-2}])] &= E[(\phi_0 + \theta_1 \varepsilon_{t-1} + \varepsilon_t - \phi_0)(\phi_0 + \theta_1 \varepsilon_{t-3} + \varepsilon_{t-2} - \phi_0)] \\
&= E[(\theta_1 \varepsilon_{t-1} + \varepsilon_t)(\theta_1 \varepsilon_{t-3} + \varepsilon_{t-2})] \\
&= E[\theta_1 \varepsilon_{t-1} \varepsilon_{t-2} + \theta_1 \varepsilon_{t-3} \varepsilon_t + \varepsilon_t \varepsilon_{t-2} + \theta_1^2 \varepsilon_{t-1} \varepsilon_{t-3}] \\
&= \theta_1 E[\varepsilon_{t-1} \varepsilon_{t-2}] + \theta_1 E[\varepsilon_{t-3} \varepsilon_t] + E[\varepsilon_t \varepsilon_{t-2}] + \theta_1^2 E[\varepsilon_{t-1} \varepsilon_{t-3}] \\
&= 0 + 0 + 0 + 0 \\
&= 0
\end{aligned} \tag{4.12}$$

By inspection of eq. (4.12) it follows that $\gamma_s = E[(y_t - E[y_t])(y_{t-s} - E[y_{t-s}])] = 0$ for $s \geq 2$.

The MA(1) can be generalized into the class of MA(Q) processes by including additional lagged errors.

Definition 4.7 (Moving Average Process of Order Q). A Moving Average process of order Q , abbreviated MA(Q), has dynamics which follow

$$y_t = \phi_0 + \sum_{q=1}^Q \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (4.13)$$

where ε_t is white noise series with the additional property that $E_{t-1}[\varepsilon_t] = 0$.

The following properties hold in higher order moving averages:

- $E[y_t] = \phi_0$
- $V[y_t] = (1 + \sum_{q=1}^Q \theta_q^2) \sigma^2$
- $E[(y_t - E[y_t])(y_{t-s} - E[y_{t-s}])] = \sigma^2 \sum_{i=0}^{Q-s} \theta_i \theta_{i+s}$ for $s \leq Q$ where $\theta_0 = 1$.
- $E[(y_t - E[y_t])(y_{t-s} - E[y_{t-s}])] = 0$ for $s > Q$

4.3.2 Autoregressive Processes

The other subclass of ARMA processes is the autoregressive process.

Definition 4.8 (First Order Autoregressive Process). A first order autoregressive process, abbreviated AR(1), has dynamics which follow

$$y_t = \phi_0 + \phi_1 y_{t-1} + \varepsilon_t \quad (4.14)$$

where ε_t is a white noise process with the additional property that $E_{t-1}[\varepsilon_t] = 0$.

Unlike the MA(1) process, y appears on both sides of the equation. However, this is only a convenience and the process can be recursively substituted to provide an expression that depends only on the errors, ε_t and an initial condition.

$$\begin{aligned} y_t &= \phi_0 + \phi_1 y_{t-1} + \varepsilon_t \\ y_t &= \phi_0 + \phi_1 (\phi_0 + \phi_1 y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\ y_t &= \phi_0 + \phi_1 \phi_0 + \phi_1^2 y_{t-2} + \varepsilon_t + \phi_1 \varepsilon_{t-1} \\ y_t &= \phi_0 + \phi_1 \phi_0 + \phi_1^2 (\phi_0 + \phi_1 y_{t-3} + \varepsilon_{t-2}) + \varepsilon_t + \phi_1 \varepsilon_{t-1} \\ y_t &= \phi_0 + \phi_1 \phi_0 + \phi_1^2 \phi_0 + \phi_1^3 y_{t-3} + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_1^2 \varepsilon_{t-2} \\ &\vdots & \vdots \\ y_t &= \sum_{i=0}^{t-1} \phi_1^i \phi_0 + \sum_{i=0}^{t-1} \phi_1^i \varepsilon_{t-i} + \phi_1^t y_0 \end{aligned}$$

Using backward substitution, an AR(1) can be expressed as an MA(t). In many cases the initial condition is unimportant and the AR process can be assumed to have begun long ago in the past.

As long as $|\phi_1| < 1$, $\lim_{t \rightarrow \infty} \phi^t y_0 \rightarrow 0$ and the effect of an initial condition will be small. Using the “infinite history” version of an AR(1), and assuming $|\phi_1| < 1$, the solution simplifies to

$$\begin{aligned} y_t &= \phi_0 + \phi_1 y_{t-1} + \varepsilon_t \\ y_t &= \sum_{i=0}^{\infty} \phi_1^i \phi_0 + \sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-i} \\ y_t &= \frac{\phi_0}{1 - \phi_1} + \sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-i} \end{aligned} \quad (4.15)$$

where the identity $\sum_{i=0}^{\infty} \phi_1^i = (1 - \phi_1)^{-1}$ is used in the final solution. This expression of an AR process is known as an MA(∞) representation and it is useful for deriving standard properties.

The unconditional mean of an AR(1) is

$$\begin{aligned} E[y_t] &= E\left[\frac{\phi_0}{1 - \phi_1} + \sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-i}\right] \\ &= \frac{\phi_0}{1 - \phi_1} + \sum_{i=0}^{\infty} \phi_1^i E[\varepsilon_{t-i}] \\ &= \frac{\phi_0}{1 - \phi_1} + \sum_{i=0}^{\infty} \phi_1^i 0 \\ &= \frac{\phi_0}{1 - \phi_1}. \end{aligned} \quad (4.16)$$

The unconditional mean can be alternatively derived noting that, as long as $\{y_t\}$ is covariance stationary, that $E[y_t] = E[y_{t-1}] = \mu$, and so

$$\begin{aligned} E[y_t] &= E[\phi_0 + \phi_1 y_{t-1} + \varepsilon_{t-1}] \\ E[y_t] &= \phi_0 + \phi_1 E[y_{t-1}] + E[\varepsilon_{t-1}] \\ \mu &= \phi_0 + \phi_1 \mu + 0 \\ \mu - \phi_1 \mu &= \phi_0 \\ \mu (1 - \phi_1) &= \phi_0 \\ E[y_t] &= \frac{\phi_0}{1 - \phi_1} \end{aligned} \quad (4.17)$$

The \mathcal{F}_{t-1} -conditional expectation is

$$\begin{aligned} E_{t-1}[y_t] &= E_{t-1}[\phi_0 + \phi_1 y_{t-1} + \varepsilon_t] \\ &= \phi_0 + \phi_1 E_{t-1}[y_{t-1}] + E_{t-1}[\varepsilon_t] \\ &= \phi_0 + \phi_1 y_{t-1} + 0 \\ &= \phi_0 + \phi_1 y_{t-1} \end{aligned} \quad (4.18)$$

since $y_{t-1} \in \mathcal{F}_{t-1}$. The unconditional and conditional variances are

$$\begin{aligned}
V[y_t] &= E[(y_t - E[y_t])^2] \\
&= E\left[\left(\frac{\phi_0}{1-\phi_1} + \sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-i} - \frac{\phi_0}{1-\phi_1}\right)^2\right] \\
&= E\left[\left(\sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-i}\right)^2\right] \\
&= E\left[\sum_{i=0}^{\infty} \phi_1^{2i} \varepsilon_{t-i}^2 + \sum_{i=0}^{\infty} \sum_{j=0, i \neq j}^{\infty} \phi_1^{i+j} \varepsilon_{t-i} \varepsilon_{t-j}\right] \\
&= E\left[\sum_{i=0}^{\infty} \phi_1^{2i} \varepsilon_{t-i}^2\right] + E\left[\sum_{i=0}^{\infty} \sum_{j=0, i \neq j}^{\infty} \phi_1^{i+j} \varepsilon_{t-i} \varepsilon_{t-j}\right] \\
&= \sum_{i=0}^{\infty} \phi_1^{2i} E[\varepsilon_{t-i}^2] + \sum_{i=0}^{\infty} \sum_{j=0, i \neq j}^{\infty} \phi_1^{i+j} E[\varepsilon_{t-i} \varepsilon_{t-j}] \\
&= \sum_{i=0}^{\infty} \phi_1^{2i} \sigma^2 + \sum_{i=0}^{\infty} \sum_{j=0, i \neq j}^{\infty} \phi_1^{i+j} 0 \\
&= \frac{\sigma^2}{1-\phi_1^2}
\end{aligned} \tag{4.19}$$

where the expression for the unconditional variance uses the identity that $\sum_{i=0}^{\infty} \phi_1^{2i} = \frac{1}{1-\phi_1^2}$ and $E[\varepsilon_{t-i} \varepsilon_{t-j}] = 0$ follows from the white noise assumption. Again, assuming covariance stationarity and so $V[y_t] = V[y_{t-1}]$, the variance can be directly computed,

$$\begin{aligned}
V[y_t] &= V[\phi_0 + \phi_1 y_{t-1} + \varepsilon_t] \\
V[y_t] &= V[\phi_0] + V[\phi_1 y_{t-1}] + V[\varepsilon_t] + 2\text{Cov}[\phi_1 y_{t-1}, \varepsilon_t] \\
V[y_t] &= 0 + \phi_1^2 V[y_{t-1}] + \sigma^2 + 2 \cdot 0 \\
V[y_t] &= \phi_1^2 V[y_t] + \sigma^2 \\
V[y_t] - \phi_1^2 V[y_t] &= \sigma^2 \\
V[y_t](1 - \phi_1^2) &= \sigma^2 \\
V[y_t] &= \frac{\sigma^2}{1 - \phi_1^2}
\end{aligned} \tag{4.20}$$

where $\text{Cov}[y_{t-1}, \varepsilon_t] = 0$ follows from the white noise assumption since y_{t-1} is a function of $\varepsilon_{t-1}, \varepsilon_{t-2}, \dots$. The conditional variance is

$$\begin{aligned}
V_{t-1}[y_t] &= E_{t-1} \left[(\phi_1 y_{t-1} + \varepsilon_t - \phi_1 y_{t-1})^2 \right] \\
&= E_{t-1} [\varepsilon_t^2] \\
&= \sigma_t^2
\end{aligned} \tag{4.21}$$

Again, the unconditional variance is uniformly larger than the average conditional variance ($E[\sigma_t^2] = \sigma^2$) and the variance explodes as $|\phi_1|$ approaches 1 or -1. Finally, the autocovariances can be derived,

$$E[(y_t - E[y_t])(y_{t-s} - E[y_{t-s}])] = E \left[\left(\frac{\phi_0}{1-\phi_1} + \sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-i} - \frac{\phi_0}{1-\phi_1} \right) \right. \tag{4.22}$$

$$\left. \times \left(\frac{\phi_0}{1-\phi_1} + \sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-s-i} - \frac{\phi_0}{1-\phi_1} \right) \right]$$

$$= E \left[\left(\sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-i} \right) \left(\sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-s-i} \right) \right]$$

$$= E \left[\left(\sum_{i=0}^{s-1} \phi_1^i \varepsilon_{t-i} + \sum_{i=s}^{\infty} \phi_1^i \varepsilon_{t-i} \right) \left(\sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-s-i} \right) \right]$$

$$= E \left[\left(\sum_{i=0}^{s-1} \phi_1^i \varepsilon_{t-i} + \sum_{i=0}^{\infty} \phi_1^s \phi_1^i \varepsilon_{t-s-i} \right) \left(\sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-s-i} \right) \right]$$

$$= E \left[\left(\sum_{i=0}^{s-1} \phi_1^i \varepsilon_{t-i} \right) \left(\sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-s-i} \right) \right.$$

$$\left. + \phi_1^s \left(\sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-s-i} \right) \left(\sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-s-i} \right) \right]$$

$$= E \left[\left(\sum_{i=0}^{s-1} \phi_1^i \varepsilon_{t-i} \right) \left(\sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-s-i} \right) \right]$$

$$+ E \left[\phi_1^s \left(\sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-s-i} \right) \left(\sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-s-i} \right) \right]$$

$$= 0 + \phi_1^s E \left[\left(\sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-s-i} \right) \left(\sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-s-i} \right) \right]$$

$$= 0 + \phi_1^s V[y_{t-s}]$$

$$= \phi_1^s \frac{\sigma^2}{1-\phi_1^2}$$

An alternative approach to deriving the autocovariance is to note that $y_t - \mu = \sum_{i=0}^{s-1} \phi_1^i \varepsilon_{t-i} + \phi^s (y_{t-s} - \mu)$ where $\mu = E[y_t] = E[y_{t-s}]$. Using this identify, the autocovariance can be derived

$$\begin{aligned}
E[(y_t - E[y_t])(y_{t-s} - E[y_{t-s}])] &= E \left[\left(\sum_{i=0}^{s-i} \phi_1^i \varepsilon_{t-i} + \phi^s (y_{t-s} - \mu) \right) (y_{t-s} - \mu) \right] \\
&= E \left[\left(\sum_{i=0}^{s-i} \phi_1^i \varepsilon_{t-i} \right) (y_{t-s} - \mu) + (\phi^s (y_{t-s} - \mu) (y_{t-s} - \mu)) \right] \\
&= E \left[\left(\sum_{i=0}^{s-i} \phi_1^i \varepsilon_{t-i} \right) (y_{t-s} - \mu) \right] + E[(\phi^s (y_{t-s} - \mu) (y_{t-s} - \mu))] \\
&= 0 + \phi^s E[(y_{t-s} - \mu) (y_{t-s} - \mu)] \\
&= \phi^s V[y_{t-s}] \\
&= \phi_1^s \frac{\sigma^2}{1 - \phi_1^2}
\end{aligned} \tag{4.26}$$

where the white noise assumption is used to ensure that $E[\varepsilon_{t-u} (y_{t-s} - \mu)] = 0$ when $u > s$.

The AR(1) can be extended to the AR(P) class by including additional lags of y_t .

Definition 4.9 (Autoregressive Process of Order P). An Autoregressive process of order P (AR(P)) has dynamics which follow

$$y_t = \phi_0 + \sum_{p=1}^P \phi_p y_{t-p} + \varepsilon_t \tag{4.27}$$

where ε_t is white noise series with the additional property that $E_{t-1}[\varepsilon_t] = 0$.

Some of the more useful properties of general AR process are:

- $E[y_t] = \frac{\phi_0}{1 - \sum_{p=1}^P \phi_p}$
- $V[y_t] = \frac{\sigma^2}{1 - \sum_{p=1}^P \phi_p \rho_p}$ where ρ_p is the p^{th} autocorrelation.
- $V[y_t]$ is infinite if $\sum_{p=1}^P \phi_p \geq 1$
- $E[(y_t - E[y_t])(y_{t-s} - E[y_{t-s}])] \neq 0$ for any s (in general, although certain parameterizations may produce some 0 autocovariances).

These four properties point to some important regularities of AR processes. First, the mean is only finite if $\sum_{p=1}^P \phi_p < 1$. Second, the autocovariances are (generally) not zero, unlike those of an MA processes ($\gamma_s = 0$ for $|s| > Q$). This difference in the behavior of the autocovariances plays an important role in model building. Explicit expressions for the variance and autocovariance of higher order AR processes can be found in appendix 4.A.

4.3.3 Autoregressive-Moving Average Processes

Putting these two processes together yields the complete class of ARMA processes.

Definition 4.10 (Autoregressive-Moving Average Process). An Autoregressive Moving Average process with orders P and Q (ARMA(P, Q)) has dynamics which follow

$$y_t = \phi_0 + \sum_{p=1}^P \phi_p y_{t-p} + \sum_{q=1}^Q \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (4.28)$$

where ε_t is a white noise process with the additional property that $E_{t-1} [\varepsilon_t] = 0$.

Again, consider the simplest ARMA(1,1) process that includes a constant term,

$$y_t = \phi_0 + \phi_1 y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t$$

To derive the properties of this model it is useful to convert the ARMA(1,1) into its infinite lag representation using recursive substitution,

$$\begin{aligned} y_t &= \phi_0 + \phi_1 y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t & (4.29) \\ y_t &= \phi_0 + \phi_1 (\phi_0 + \phi_1 y_{t-2} + \theta_1 \varepsilon_{t-2} + \varepsilon_{t-1}) + \theta_1 \varepsilon_{t-1} + \varepsilon_t \\ y_t &= \phi_0 + \phi_1 \phi_0 + \phi_1^2 y_{t-2} + \phi_1 \theta_1 \varepsilon_{t-2} + \phi_1 \varepsilon_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t \\ y_t &= \phi_0 + \phi_1 \phi_0 + \phi_1^2 (\phi_0 + \phi_1 y_{t-3} + \theta_1 \varepsilon_{t-3} + \varepsilon_{t-2}) + \phi_1 \theta_1 \varepsilon_{t-2} + \phi_1 \varepsilon_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t \\ y_t &= \phi_0 + \phi_1 \phi_0 + \phi_1^2 \phi_0 + \phi_1^3 y_{t-3} + \phi_1^2 \theta_1 \varepsilon_{t-3} + \phi_1^2 \varepsilon_{t-2} + \phi_1 \theta_1 \varepsilon_{t-2} + \phi_1 \varepsilon_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t \\ &\vdots & \vdots \\ y_t &= \sum_{i=0}^{\infty} \phi_1^i \phi_0 + \varepsilon_t + \sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \varepsilon_{t-i-1} \\ y_t &= \frac{\phi_0}{1 - \phi_1} + \varepsilon_t + \sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \varepsilon_{t-i-1}. \end{aligned}$$

Using the infinite lag representation, the unconditional and conditional means can be computed,

$$\begin{aligned} E[y_t] &= E \left[\frac{\phi_0}{1 - \phi_1} + \varepsilon_t + \sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \varepsilon_{t-i-1} \right] & (4.30) \\ &= \frac{\phi_0}{1 - \phi_1} + E[\varepsilon_t] + \sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) E[\varepsilon_{t-i-1}] \\ &= \frac{\phi_0}{1 - \phi_1} + 0 + \sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) 0 \\ &= \frac{\phi_0}{1 - \phi_1} \end{aligned}$$

and

$$\begin{aligned} E_{t-1}[y_t] &= E_{t-1}[\phi_0 + \phi_1 y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t] \\ &= \phi_0 + \phi_1 E_{t-1}[y_{t-1}] + \theta_1 E_{t-1}[\varepsilon_{t-1}] + E_{t-1}[\varepsilon_t] \\ &= \phi_0 + \phi_1 y_{t-1} + \theta_1 \varepsilon_{t-1} + 0 \\ &= \phi_0 + \phi_1 y_{t-1} + \theta_1 \varepsilon_{t-1} \end{aligned} \quad (4.31)$$

Since y_{t-1} and ε_{t-1} are in the time- $t-1$ information set, these variables pass through the conditional expectation. The unconditional variance can be tediously derived (see appendix 4.A.3 for the complete derivation)

$$V[y_t] = \sigma^2 \left(\frac{1 + 2\phi_1\theta_1 + \theta_1^2}{1 - \phi_1^2} \right) \quad (4.32)$$

The conditional variance is identical to that in the AR(1) or MA(1), $V_{t-1}[y_t] = \sigma_t^2$, and, if ε_t is homoskedastic, $V_{t-1}[y_t] = \sigma^2$.

The unconditional mean of an ARMA is the same as an AR since the moving average terms, which are all mean zero, make no contribution. The variance is more complicated, and it may be larger or smaller than an AR(1) with the same autoregressive parameter (ϕ_1). The variance will only be smaller if ϕ_1 and θ_1 have opposite signs and $2\phi_1\theta_1 < \theta_1^2$. Deriving the autocovariance is straightforward but tedious and is presented in appendix 4.A.

4.4 Difference Equations

Before turning to the analysis of the stationarity conditions for ARMA processes, it is useful to develop an understanding of the stability conditions in a setting without random shocks.

Definition 4.11 (Linear Difference Equation). An equation of the form

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_P y_{t-P} + x_t. \quad (4.33)$$

is known as a Pth order linear difference equation where the series $\{x_t\}$ is known as the driving process.

Linear difference equation nest ARMA processes which can be seen by setting x_t equal to the shock plus the moving average component of the ARMA process,

$$x_t = \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_Q \varepsilon_{t-Q} + \varepsilon_t.$$

Stability conditions depend crucially on the solution to the linear difference equation.

Definition 4.12 (Solution). A solution to a linear difference equation expresses the linear difference equation

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_P y_{t-P} + x_t. \quad (4.34)$$

as a function of only $\{x_i\}_{i=1}^t$, a constant and, when y_t has finite history, an initial value y_0 .

Consider a first order linear difference equation

$$y_t = \phi_0 + \phi_1 y_{t-1} + x_t.$$

Starting from an initial value y_0 ,

$$y_1 = \phi_0 + \phi_1 y_0 + x_1$$

$$\begin{aligned} y_2 &= \phi_0 + \phi_1(\phi_0 + \phi_1 y_0 + x_1) + x_2 \\ &= \phi_0 + \phi_1 \phi_0 + \phi_1^2 y_0 + x_2 + \phi_1 x_1 \end{aligned}$$

$$\begin{aligned} y_3 &= \phi_0 + \phi_1 y_2 + x_2 \\ &= \phi_0 + \phi_1(\phi_0 + \phi_1 \phi_0 + \phi_1^2 y_0 + \phi_1 x_1 + x_2) + x_2 \\ &= \phi_0 + \phi_1 \phi_0 + \phi_1^2 \phi_0 + \phi_1^3 y_0 + x_3 + \phi_1 x_2 + \phi_1^2 x_1 \end{aligned}$$

Continuing these iterations, a pattern emerges:

$$y_t = \phi_1^t y_0 + \sum_{i=0}^{t-1} \phi_1^i \phi_0 + \sum_{i=0}^{t-1} \phi_1^i x_{t-i} \quad (4.35)$$

This is a solution since it expresses y_t as a function of only $\{x_t\}$, y_0 and constants. When no initial condition is given (or the series is assumed to be infinite), the solution can be found by solving backward

$$y_t = \phi_0 + \phi_1 y_{t-1} + x_t$$

$$\begin{aligned} y_{t-1} &= \phi_0 + \phi_1 y_{t-2} + x_{t-1} \Rightarrow \\ y_t &= \phi_0 + \phi_1(\phi_0 + \phi_1 y_{t-2} + x_{t-1}) + x_t \\ &= \phi_0 + \phi_1 \phi_0 + \phi_1^2 y_{t-2} + x_t + \phi_1 x_{t-1} \\ y_{t-2} &= \phi_0 + \phi_1 y_{t-3} + x_{t-2} \Rightarrow \\ y_t &= \phi_0 + \phi_1 \phi_0 + \phi_1^2 (\phi_0 + \phi_1 y_{t-3} + x_{t-2}) + x_t + \phi_1 x_{t-1} \\ &= \phi_0 + \phi_1 \phi_0 + \phi_1^2 \phi_0 + \phi_1^3 y_{t-3} + x_t + \phi_1 x_{t-1} + \phi_1^2 x_{t-2} \end{aligned}$$

which leads to the approximate solution

$$y_t = \sum_{i=0}^{s-1} \phi_1^i \phi_0 + \sum_{i=0}^{s-1} \phi_1^i x_{t-i} + \phi_1^s y_{t-s}.$$

To understand the behavior of this solution, it is necessary to take limits. If $|\phi_1| < 1$, $\lim_{s \rightarrow \infty} \phi_1^s y_{t-s}$ goes to zero (as long as y_{t-s} is bounded) and the solution simplifies to

$$y_t = \phi_0 \sum_{i=0}^{\infty} \phi_1^i + \sum_{i=0}^{\infty} \phi_1^i x_{t-i}. \quad (4.36)$$

Noting that, as long as $|\phi_1| < 1$, $\sum_{i=0}^{\infty} \phi_1^i = 1/(1 - \phi_1)$,

$$y_t = \frac{\phi_0}{1 - \phi_1} + \sum_{i=0}^{\infty} \phi_1^i x_{t-i} \quad (4.37)$$

is the solution to this problem with an infinite history. The solution concept is important because it clarifies the relationship between observations in the distant past and the current observation, and if $\lim_{s \rightarrow \infty} \phi_1^s y_{t-s}$ does not converge to zero then observations arbitrarily far in the past have an influence on the value of y today.

When $|\phi_1| > 1$ then this system is said to be *nonconvergent* since ϕ_1^t diverges as t grows large and values in the past are not only important, they will dominate when determining the current value. In the special case where $\phi_1 = 1$,

$$y_t = \phi_0 t + \sum_{i=0}^{\infty} x_{t-i},$$

which is a random walk when $\{x_t\}$ is a white noise process, and the influence of a single x_t never diminishes. Direct substitution can be used to find the solution of higher order linear difference equations at the cost of more tedium. A simpler alternative is to focus on the core component of a linear difference equation, the linear homogeneous equation.

4.4.1 Homogeneous Difference Equations

When the number of lags grows large (3 or greater), solving linear difference equations by substitution is tedious. The key to understanding linear difference equations is the study of the homogeneous portion of the equation. In the general linear difference equation,

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_P y_{t-P} + x_t$$

the homogenous portion is defined as the terms involving only y ,

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_P y_{t-P}. \quad (4.38)$$

The intuition behind studying this portion of the system is that given the sequence of $\{x_t\}$, all of the dynamics and the stability of the system are determined by the relationship between contemporaneous y_t and its lagged values which allows the determination of the parameter values where the system is stable. Again, consider the homogeneous portions of the simple 1st order system,

$$y_t = \phi_1 y_{t-1} + x_t$$

which has the homogeneous portion

$$y_t = \phi_1 y_{t-1}.$$

To find solutions to this equation, one can try trial and error: one obvious solution is 0 since $0 = \phi \cdot 0$. It is easy to show that

$$y_t = \phi_1^t y_0$$

is also a solution by examining the solution to the linear difference equation in eq. (4.35). But then so is any solution of the form $c\phi_1^t$ for an arbitrary constant c . How?

$$\begin{aligned} y_t &= c\phi_1^t \\ y_{t-1} &= c\phi_1^{t-1} \end{aligned}$$

and

$$y_t = \phi_1 y_{t-1}$$

Putting these two together shows that

$$\begin{aligned} y_t &= \phi_1 y_{t-1} \\ c\phi_1^t &= \phi_1 y_{t-1} \\ c\phi_1^t &= \phi_1 c\phi_1^{t-1} \\ c\phi_1^t &= c\phi_1^t \end{aligned}$$

and there are many solutions. However, from these it is possible to discern when the solution will converge to zero and when it will explode:

- If $|\phi_1| < 1$ the system converges to 0. If ϕ_1 is also negative, the solution oscillates, while if ϕ_1 is greater than 0, the solution decays exponentially.
- If $|\phi_1| > 1$ the system diverges, again oscillating if negative and growing exponentially if positive.
- If $\phi_1 = 1$, the system is stable and all values are solutions. For example $1 = 1 \cdot 1$, $2 = 1 \cdot 2$, etc.
- If $\phi_1 = -1$, the system is *metastable*. The values, in absolute terms, are unchanged, but it oscillates between + and -.

These categories will play important roles in examining the dynamics of larger equations since they determine how past shocks will affect current values of y_t . When the order is greater than 1, there is an easier approach to examining the stability of the system. Consider the second order linear difference system,

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + x_t$$

and again focus on the homogeneous portion,

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2}.$$

This equation can be rewritten

$$y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} = 0$$

so any solution of the form

$$\begin{aligned} cz^t - \phi_1 cz^{t-1} - \phi_2 cz^{t-2} &= 0 \\ cz^{t-2} (z^2 - \phi_1 z - \phi_2) &= 0 \end{aligned} \tag{4.39}$$

will solve this equation.⁴ Dividing through by cz^{t-2} , this is equivalent to

$$z^2 - \phi_1 z - \phi_2 = 0 \tag{4.40}$$

and the solutions to this quadratic polynomial are given by the quadratic formula,

$$c_1, c_2 = \frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{2} \tag{4.41}$$

The roots of the equation, c_1 and c_2 , play the same role as ϕ_1 in the 1st order case.⁵ If $|c_1| < 1$ and $|c_2| < 1$, the system is convergent. With two roots both smaller than 1 there are three interesting cases:

Case 1: Both roots are real and positive. In this case, the system will exponentially dampen and not oscillate.

Case 2: Both roots are imaginary (of the form $c + di$ where $i = \sqrt{-1}$) and distinct, or real and at least one negative. In this case, the absolute value of the roots (also called the modulus, defined as $\sqrt{c^2 + d^2}$ for an imaginary number $c + di$) is less than 1, and so the system will be convergent but oscillate.

Case 3: Real but the same. This occurs when $\phi_1^2 + 4\phi_2 = 0$. Since there is only one root, the system is convergent if it is less than 1 in absolute value, which requires that $|\phi_1| < 2$.

If either are greater than 1 in absolute terms, the system is divergent.

4.4.2 Lag Operators

Before proceeding to higher order models, it is necessary to define the lag operator. Lag operations are a particularly useful tool in the analysis of time series and are nearly self-descriptive.⁶

Definition 4.13 (Lag Operator). The lag operator is denoted L and is defined as the operator that has

⁴The solution can only be defined up to a constant, c , since the right hand side is 0. Thus, multiplying both by a constant, the solution will still be valid.

⁵In the first order case, $y_t = \phi_1 y_{t-1}$, so $y_t - \phi_1 y_{t-1} = 0$. The solution has the property that $z^t - \phi_1 z^{t-1} = 0$ so $z - \phi_1 = 0$, which has the single solution $c = \phi_1$.

⁶In some texts, the lag operator is known as the backshift operator, and L is replaced with B .

the following properties:

$$\begin{aligned} Ly_t &= y_{t-1} \\ L^2y_t &= y_{t-2} \\ L^i y_t &= y_{t-i} \\ L(L(y_t)) &= L(y_{t-1}) = y_{t-2} = L^2 y_t \\ (1 - L - L^2)y_t &= y_t - Ly_t - L^2 y_t = y_t - y_{t-1} - y_{t-2} \end{aligned}$$

The last equation above is particularly useful when studying autoregressive processes. One additional property of the lag operator is that the lag of a constant is just the constant, i.e. $Lc = c$.

4.4.3 Higher Order Linear Homogenous Equations

Stability analysis can be applied to higher order systems by forming the characteristic equation and finding the characteristic roots.

Definition 4.14 (Characteristic Equation). Let y_t follow a P^{th} order linear difference equation

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_P y_{t-P} + x_t \quad (4.42)$$

which can be rewritten as

$$\begin{aligned} y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \dots - \phi_P y_{t-P} &= \phi_0 + x_t \\ (1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_P L^P) y_t &= \phi_0 + x_t \end{aligned} \quad (4.43)$$

The characteristic equation of this process is

$$z^P - \phi_1 z^{P-1} - \phi_2 z^{P-2} - \dots - \phi_{P-1} z - \phi_P = 0 \quad (4.44)$$

The characteristic roots are the solutions to this equation and most econometric packages will return the roots of the characteristic polynomial when an ARMA model is estimated.

Definition 4.15 (Characteristic Root). Let

$$z^P - \phi_1 z^{P-1} - \phi_2 z^{P-2} - \dots - \phi_{P-1} z - \phi_P = 0 \quad (4.45)$$

be the characteristic polynomial associated with a P^{th} order linear difference equation. The P characteristic roots, c_1, c_2, \dots, c_P are defined as the solution to this polynomial

$$(z - c_1)(z - c_2) \dots (z - c_P) = 0 \quad (4.46)$$

The conditions for stability are the same for higher order systems as they were for first and second order systems: all roots c_p , $p = 1, 2, \dots, P$ must satisfy $|c_p| < 1$ (again, if complex, $|\cdot|$ means modulus). If any $|c_p| > 1$ the system is divergent. If one or more $|c_p| = 1$ and none are larger, the system will exhibit unit root (random walk) behavior.

These results are the key to understanding important properties of linear time-series models which turn out to be *stationary* if the corresponding linear homogeneous system is convergent, i.e. $|c_p| < 1$, $p = 1, 2, \dots, P$.

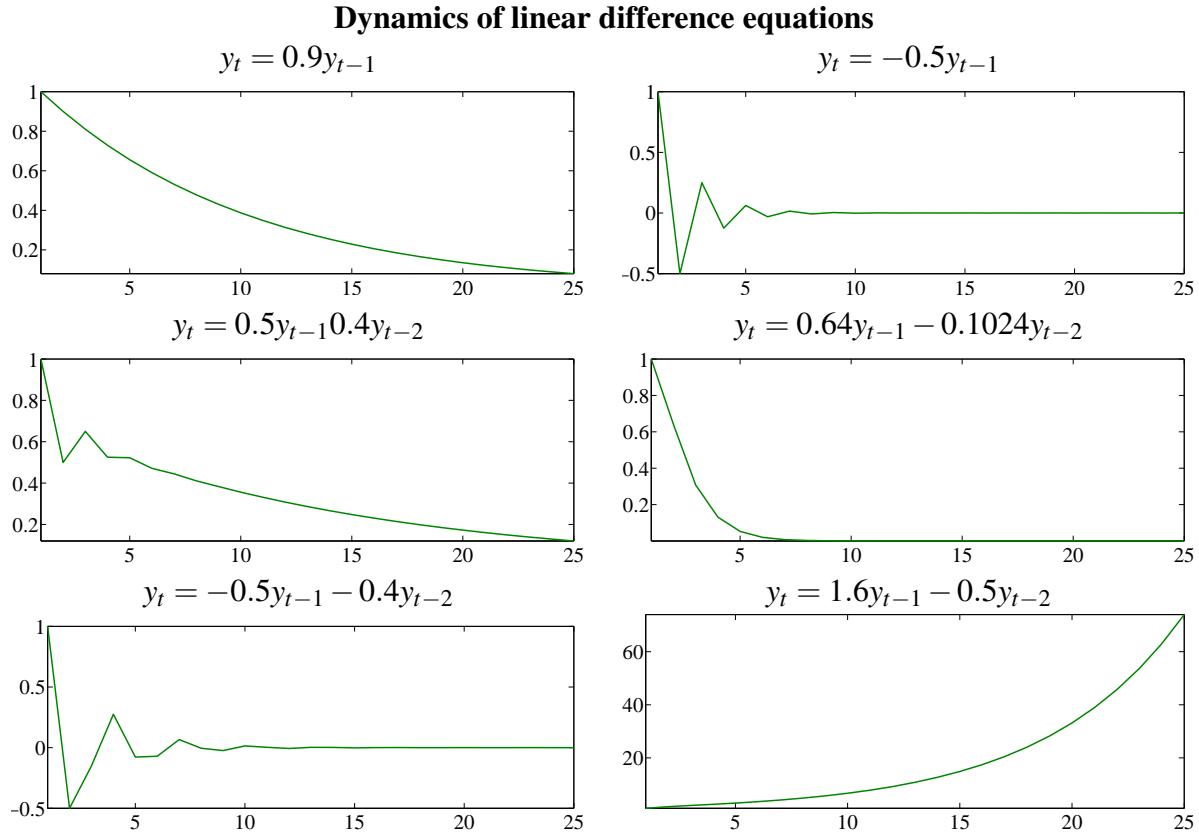


Figure 4.1: These six plots correspond to the dynamics of the six linear homogeneous systems described in the text. All processes received a unit shock at $t = 1$ ($x_1 = 1$) and no other shocks ($x_j = 0$, $j \neq 1$). Pay close attention to the roots of the characteristic polynomial and the behavior of the system (exponential decay, oscillation and/or explosion).

4.4.4 Example: Characteristic Roots and Stability

Consider 6 linear difference equations, their characteristic equation, and the roots:

- $y_t = 0.9y_{t-1} + x_t$
 - Characteristic Equation: $z-0.9=0$
 - Characteristic Root: $z=0.9$
- $y_t = -0.5y_{t-1} + x_t$
 - Characteristic Equation: $z+0.5=0$
 - Characteristic Root: $z=-0.5$
- $y_t = 0.5y_{t-1} + 0.4y_{t-2} + x_t$
 - Characteristic Equation: $z^2 - 0.5z - 0.4 = 0$

- Characteristic Roots: $z = 0.93, -0.43$
- $y_t = 0.64y_{t-1} - 0.1024y_{t-2} + x_t$
 - Characteristic Equation: $z^2 - 0.64z + 0.1024 = 0$
 - Characteristic Roots: $z = 0.32, 0.32$ (identical)
- $y_t = -0.5y_{t-1} - 0.4y_{t-2} + x_t$
 - Characteristic Equation: $z^2 + 0.5z + 0.4 = 0$
 - Characteristic Roots (Modulus): $z = -0.25 + 0.58i(0.63), -0.25 - 0.58i(0.63)$
- $y_t = 1.6y_{t-1} - 0.5y_{t-2} + x_t$
 - Characteristic Equation: $z^2 - 1.6z + 0.5 = 0$
 - Characteristic Roots: $z = 1.17, 0.42$

The plots in figure 4.1 show the effect of a unit (1) shock at $t = 1$ to the 6 linear difference systems above (all other shocks are 0). The value of the root makes a dramatic difference in the observed behavior of the series.

4.4.5 Stationarity of ARMA models

Stationarity conditions for ARMA processes can be determined using the results for the convergence of linear difference equations. First, note that any ARMA process can be written using a lag polynomial

$$\begin{aligned} y_t &= \phi_0 + \phi_1 y_{t-1} + \dots + \phi_P y_{t-P} + \theta_1 \varepsilon_{t-1} + \dots + \theta_Q \varepsilon_{t-Q} + \varepsilon_t \\ y_t - \phi_1 y_{t-1} - \dots - \phi_P y_{t-P} &= \phi_0 + \theta_1 \varepsilon_{t-1} + \dots + \theta_Q \varepsilon_{t-Q} + \varepsilon_t \\ (1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_P L^P) y_t &= \phi_0 + (1 + \theta_1 L + \theta_2 L^2 + \dots + \theta_Q L^Q) \varepsilon_t \end{aligned}$$

This is a linear difference equation, and the stability conditions depend on the roots of the characteristic polynomial

$$z^P - \phi_1 z^{P-1} - \phi_2 z^{P-2} - \dots - \phi_{P-1} z - \phi_P$$

An ARMA process driven by a white noise shock will be covariance stationary as long as the characteristic roots are less than one in modulus. In the simple AR(1) case, this corresponds to $|z_1| < 1$. In the AR(2) case, the region is triangular with a curved bottom and corresponds to the points $(z_1, z_2) = (-2, -1), (1, 0), (2, -2)$ (see figure 4.2). For higher order models, stability must be checked by numerically solving the characteristic equation.

The other particularly interesting point is that *all* MA processes driven by covariance stationary shocks are stationary since the homogeneous portions of an MA process has no root and thus cannot diverge.

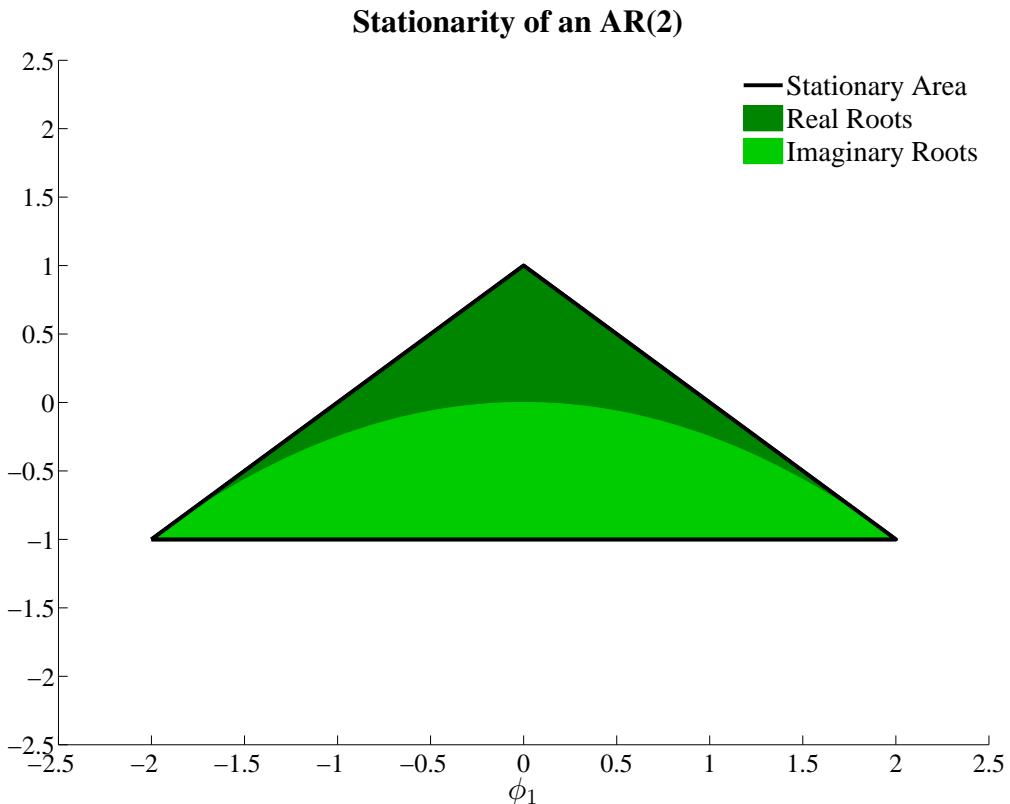


Figure 4.2: The triangular region corresponds to the values of the parameters in the AR(2) $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$. The dark region corresponds to real roots and the light region corresponds to imaginary roots.

4.5 Data and Initial Estimates

Two series will be used throughout the stationary time-series analysis section: returns on the value weighted market and the spread between the average interest rates on portfolios of Aaa-rated and Baa-rated corporate bonds, commonly known as the default spread or default premium. The VWM returns were taken from CRSP and are available from January 1927 through July 2008 and the bond yields are available from Moody's via FRED II and are available from January 1919 until July 2008. Both series are monthly.

Figure 4.3 contains plots of the two series. Table 4.1 contains parameter estimates for an AR(1), an MA(1) and an ARMA(1,1) for each series. The default spread exhibits a large autoregressive coefficient (.97) that is highly significant, but it also contains a significant moving average term and in an ARMA(1,1) both parameters are significant. The market portfolio exhibits some evidence of predictability although it is much less persistent than the default spread.⁷

⁷For information on estimating an ARMA in MATLAB, see the MATLAB supplement to this course.

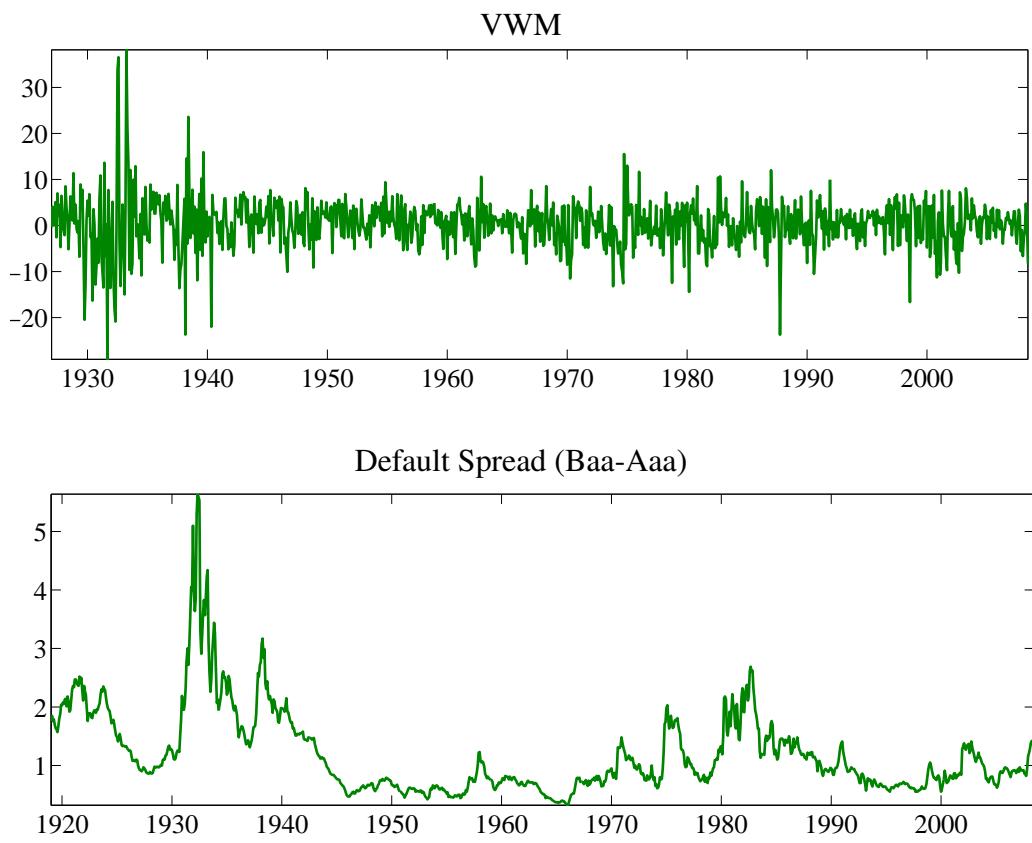


Figure 4.3: Plots of the returns on the VWM and the default spread, the spread between the yield of a portfolio of Baa-rated bonds and the yield of a portfolio of Aaa-rated bonds.

VWM				Baa-Aaa			
$\hat{\phi}_0$	$\hat{\phi}_1$	$\hat{\theta}_1$	$\hat{\sigma}$	$\hat{\phi}_0$	$\hat{\phi}_1$	$\hat{\theta}_1$	$\hat{\sigma}$
0.284 (0.108)	0.115 (0.052)		5.415	0.026 (0.284)	0.978 (0.000)		0.149
0.320 (0.096)		0.115 (0.042)	5.415	1.189 (0.000)		0.897 (0.000)	0.400
0.308 (0.137)	0.039 (0.870)	0.077 (0.724)	5.417	0.036 (0.209)	0.969 (0.000)	0.202 (0.004)	0.146

Table 4.1: Parameter estimates and p-values from an AR(1), MA(1) and ARMA(1,1) for the VWM and Baa-Aaa spread.

4.6 Autocorrelations and Partial Autocorrelations

Autoregressive processes, moving average processes and ARMA processes all exhibit different patterns in their autocorrelations and partial autocorrelations. These differences can be exploited to select a parsimonious model from the general class of ARMA processes.

4.6.1 Autocorrelations and the Autocorrelation Function

Autocorrelations are to autocovariances as correlations are to covariances. That is, the s^{th} autocorrelation is the s^{th} autocovariance divided by the product of the variance of y_t and y_{t-s} , and when a process is covariance stationary, $V[y_t] = V[y_{t-s}]$, and so $\sqrt{V[y_t]V[y_{t-s}]} = V[y_t]$.

Definition 4.16 (Autocorrelation). The autocorrelation of a covariance stationary scalar process is defined

$$\rho_s = \frac{\gamma_s}{\gamma_0} = \frac{E[(y_t - E[y_t])(y_{t-s} - E[y_{t-s}])]}{V[y_t]} \quad (4.47)$$

where γ_s is the s^{th} autocovariance.

The autocorrelation function (ACF) relates the lag length (s) and the parameters of the model to the autocorrelation.

Definition 4.17 (Autocorrelation Function). The autocorrelation function (ACF), $\rho(s)$, is a function of the population parameters that defines the relationship between the autocorrelations of a process and lag length.

The variance of a covariance stationary AR(1) is $\sigma^2(1 - \phi_1^2)^{-1}$ and the s^{th} autocovariance is $\phi^s \sigma^2(1 - \phi_1^2)^{-1}$, and so the ACF is

$$\rho(s) = \frac{\phi^s \sigma^2(1 - \phi^2)^{-1}}{\sigma^2(1 - \phi^2)^{-1}} = \phi^s. \quad (4.48)$$

Deriving ACFs of ARMA processes is a straightforward, albeit tedious, task. Further details on the derivation of the ACF of a stationary ARMA processes are presented in appendix 4.A.

4.6.2 Partial Autocorrelations and the Partial Autocorrelation Function

Partial autocorrelations are similar to autocorrelations with one important difference: the s^{th} partial autocorrelation still relates y_t and y_{t-s} but it eliminates the effects of $y_{t-1}, y_{t-2}, \dots, y_{t-(s-1)}$.

Definition 4.18 (Partial Autocorrelation). The s^{th} partial autocorrelation (ϕ_s) is defined as the population value of the regression coefficient on ϕ_s in

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_{s-1} y_{t-(s-1)} + \phi_s y_{t-s} + \varepsilon_t.$$

Like the autocorrelation function, the partial autocorrelation function (PACF) relates the partial autocorrelation to population parameters and lag length.

Definition 4.19 (Partial Autocorrelation Function). The partial autocorrelation function (PACF), $\phi(s)$, defines the relationship between the partial autocorrelations of a process and lag length. The PACF is denoted.

The partial autocorrelations are directly interpretable as population regression coefficients. The s^{th} partial autocorrelations can be computed using $s+1$ autocorrelations. Recall that the population values of $\phi_1, \phi_2, \dots, \phi_s$ in

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_{s-1} y_{t-(s-1)} + \phi_s y_{t-s} + \varepsilon_t$$

can be defined in terms of the covariance between $y_t, y_{t-1}, y_{t-2}, \dots, y_{t-s}$. Let Γ denote this covariance matrix,

$$\Gamma = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \gamma_3 & \dots & \gamma_{s-1} & \gamma_s \\ \gamma_1 & \gamma_0 & \gamma_1 & \gamma_2 & \dots & \gamma_{s-2} & \gamma_{s-1} \\ \gamma_2 & \gamma_1 & \gamma_0 & \gamma_1 & \dots & \gamma_{s-3} & \gamma_{s-2} \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ \gamma_{s-1} & \gamma_{s-2} & \gamma_{s-3} & \gamma_{s-4} & \dots & \gamma_0 & \gamma_1 \\ \gamma_s & \gamma_{s-1} & \gamma_{s-2} & \gamma_{s-3} & \dots & \gamma_1 & \gamma_0 \end{bmatrix}$$

The matrix Γ is known as a Toeplitz matrix which reflects the special symmetry it exhibits which follows from stationarity, and so $E[(y_t - \mu)(y_{t-s} - \mu)] = \gamma_s = \gamma_{-s} = E[(y_t - \mu)(y_{t+s} - \mu)]$. Γ can be decomposed in terms of γ_0 (the long-run variance) and the matrix of autocorrelations,

$$\Gamma = \gamma_0 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \dots & \rho_{s-1} & \rho_s \\ \rho_1 & 1 & \rho_1 & \rho_2 & \dots & \rho_{s-2} & \rho_{s-1} \\ \rho_2 & \rho_1 & 1 & \rho_1 & \dots & \rho_{s-3} & \rho_{s-2} \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ \rho_{s-1} & \rho_{s-2} & \rho_{s-3} & \rho_{s-1} & \dots & 1 & \rho_1 \\ \rho_s & \rho_{s-1} & \rho_{s-2} & \rho_{s-3} & \dots & \rho_1 & 1 \end{bmatrix}$$

directly by applying the definition of an autocorrelation. The population regression parameters can be computed by partitioning Γ into four blocks, γ_0 , the long-run variance of y_t , $\Gamma_{01} = \Gamma'_{10}$, the vector of covariances between y_t and $y_{t-1}, y_{t-2}, \dots, y_{t-s}$, and Γ_{11} , the covariance matrix of $y_{t-1}, y_{t-2}, \dots, y_{t-s}$.

$$\Gamma = \begin{bmatrix} \gamma_0 & \Gamma_{01} \\ \Gamma_{10} & \Gamma_{11} \end{bmatrix} = \gamma_0 \begin{bmatrix} 1 & \mathbf{R}_{01} \\ \mathbf{R}_{10} & \mathbf{R}_{11} \end{bmatrix}$$

where \mathbf{R} are vectors or matrices of autocorrelations. Using this formulation, the population regression parameters $\phi = [\phi_1, \phi_2, \dots, \phi_s]'$ are defined as

$$\phi = \Gamma_{11}^{-1} \Gamma_{10} = \gamma_0^{-1} \mathbf{R}_{11}^{-1} \gamma_0 \mathbf{R}_{10} = \mathbf{R}_{11}^{-1} \mathbf{R}_{10}. \quad (4.49)$$

The s^{th} partial autocorrelation (φ_s) is the s^{th} element in ϕ (when Γ is s by s), $\mathbf{e}_s' \mathbf{R}_{11}^{-1} \mathbf{R}_{10}$ where \mathbf{e}_s is a s by 1 vector of zeros with one in the s^{th} position.

For example, in a stationary AR(1) model, $y_t = \phi_1 y_{t-1} + \varepsilon_t$, the PACF is

$$\begin{aligned} \varphi(s) &= \phi_1^{|s|} & s = 0, 1, -1 \\ &= 0 & \text{otherwise} \end{aligned}$$

That $\varphi_0 = \phi^0 = 1$ is obvious: the correlation of a variable with itself is 1. The first partial autocorrelation is defined as the population parameter of ϕ_1 in the regression $y_t = \phi_0 + \phi_1 y_{t-1} + \varepsilon_t$. Since

the data generating process is an AR(1), $\phi_1 = \phi_1$, the autoregressive parameter. The second partial autocorrelation is defined as the population value of ϕ_2 in the regression

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t.$$

Since the DGP is an AR(1), once y_{t-1} is included, y_{t-2} has no effect on y_t and the population value of both ϕ_2 and the second partial autocorrelation, ϕ_2 , is 0. This argument holds for any higher order partial autocorrelation.

Note that the first partial autocorrelation and the first autocorrelation are both ϕ_1 in

$$y_t = \phi_0 + \phi_1 y_{t-1} + \varepsilon_t,$$

and at the second (and higher) lag these differ. The autocorrelation at $s = 2$ is the population value of ϕ_2 in the regression

$$y_t = \phi_0 + \phi_2 y_{t-2} + \varepsilon$$

while the second partial autocorrelation is the population value of from ϕ_2 in the regression

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon.$$

If the DGP were an AR(1), the second autocorrelation would be $\rho_2 = \phi_1^2$ while the second partial autocorrelation would be $\phi_2 = 0$.

4.6.2.1 Examples of ACFs and PACFs

The key to understanding the value of ACFs and PACFs lies in the distinct behavior the autocorrelations and partial autocorrelations of AR and MA processes exhibit.

- AR(P)
 - ACF dies exponentially (may oscillate, referred to as sinusoidally)
 - PACF is zero beyond P

- MA(Q)
 - ACF is zero beyond Q
 - PACF dies exponentially (may oscillate, referred to as sinusoidally)

Table 4.2 provides a summary of the ACF and PACF behavior of ARMA models and this difference forms the basis of the Box-Jenkins model selection strategy.

Process	ACF	PACF
White Noise	All 0	All 0
AR(1)	$\rho_s = \phi^s$	0 beyond lag 2
AR(P)	Decays toward zero exponentially	Non-zero through lag P, 0 thereafter
MA(1)	$\rho_1 \neq 0, \rho_s = 0, s > 1$	Decays toward zero exponentially
MA(Q)	Non-zero through lag Q, 0 thereafter	Decays toward zero exponentially
ARMA(P,Q)	Exponential Decay	Exponential Decay

Table 4.2: Behavior that the ACF and PACF for various members of the ARMA family.

4.6.3 Sample Autocorrelations and Partial Autocorrelations

Sample autocorrelations are computed using sample analogues of the population moments in the definition of an autocorrelation. Define $y_t^* = y_t - \bar{y}$ to be the demeaned series where $\bar{y} = T^{-1} \sum_{t=1}^T y_t$. The s^{th} sample autocorrelation is defined

$$\hat{\rho}_s = \frac{\sum_{t=s+1}^T y_t^* y_{t-s}^*}{\sum_{t=1}^T (y_t^*)^2} \quad (4.50)$$

although in small samples one the the corrected versions

$$\hat{\rho}_s = \frac{\frac{\sum_{t=s+1}^T y_t^* y_{t-s}^*}{T-S}}{\frac{\sum_{t=1}^T (y_t^*)^2}{T}} \quad (4.51)$$

or

$$\hat{\rho}_s = \frac{\sum_{t=s+1}^T y_t^* y_{t-s}^*}{\sqrt{\sum_{t=s+1}^T (y_t^*)^2 \sum_{t=1}^{T-s} (y_t^*)^2}}. \quad (4.52)$$

may be more accurate.

Definition 4.20 (Sample Autocorrelogram). A plot of the sample autocorrelations against the lag index in known as a sample autocorrelogram.

Inference on estimated autocorrelation coefficients depends on the null hypothesis tested and whether the data are homoskedastic. The most common assumptions are that the data are homoskedastic and that *all* of the autocorrelations are zero. In other words, $y_t - E[y_t]$ is white noise process. Under the null $H_0 : \rho_s = 0, s \neq 0$, inference can be made noting that $V[\hat{\rho}_s] = T^{-1}$ using a standard t -test,

$$\frac{\hat{\rho}_s}{\sqrt{V[\hat{\rho}_s]}} = \frac{\hat{\rho}_s}{\sqrt{T^{-1}}} = T^{\frac{1}{2}} \hat{\rho}_s \xrightarrow{d} N(0, 1). \quad (4.53)$$

A alternative null hypothesis is that the autocorrelations on lags s and above are zero but that the autocorrelations on lags $1, 2, \dots, s-1$ are unrestricted, $H_0 : \rho_j = 0, j \geq s$. Under this null, and again assuming homoskedasticity,

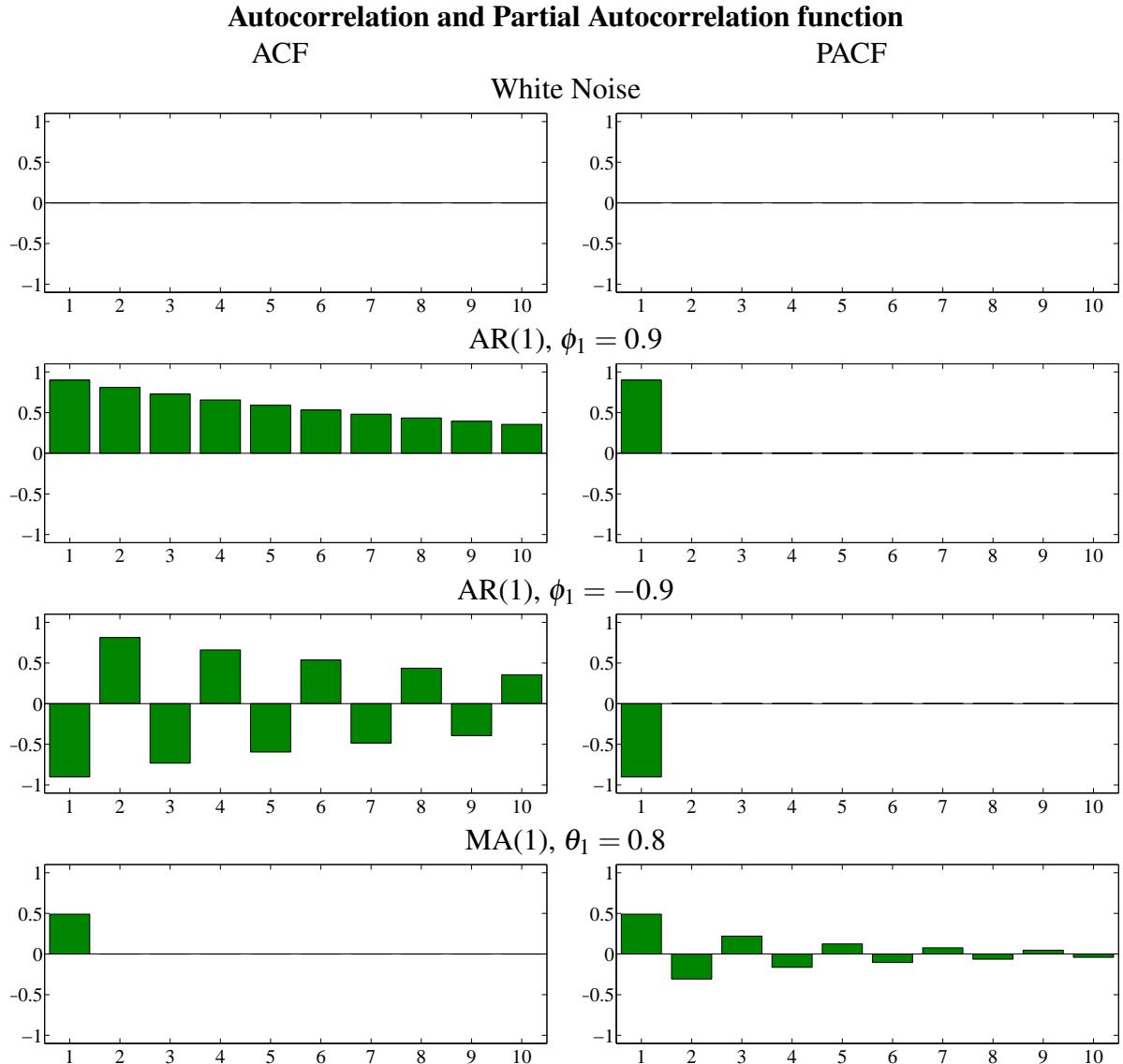


Figure 4.4: Autocorrelation function and partial autocorrelation function for 4 processes. Note the difference between how the ACF and PACF respond in AR and MA models.

$$\begin{aligned}
 \mathbf{V}[\hat{\rho}_s] &= T^{-1} && \text{for } s = 1 \\
 &= T^{-1} \left(1 + 2 \sum_{j=1}^{s-1} \hat{\rho}_j^2 \right) && \text{for } s > 1
 \end{aligned} \tag{4.54}$$

If the null is $H_0 : \rho_s = 0$ with no further restrictions on the other autocorrelations, the variance of the s^{th} autocorrelation is (assuming homoskedasticity)

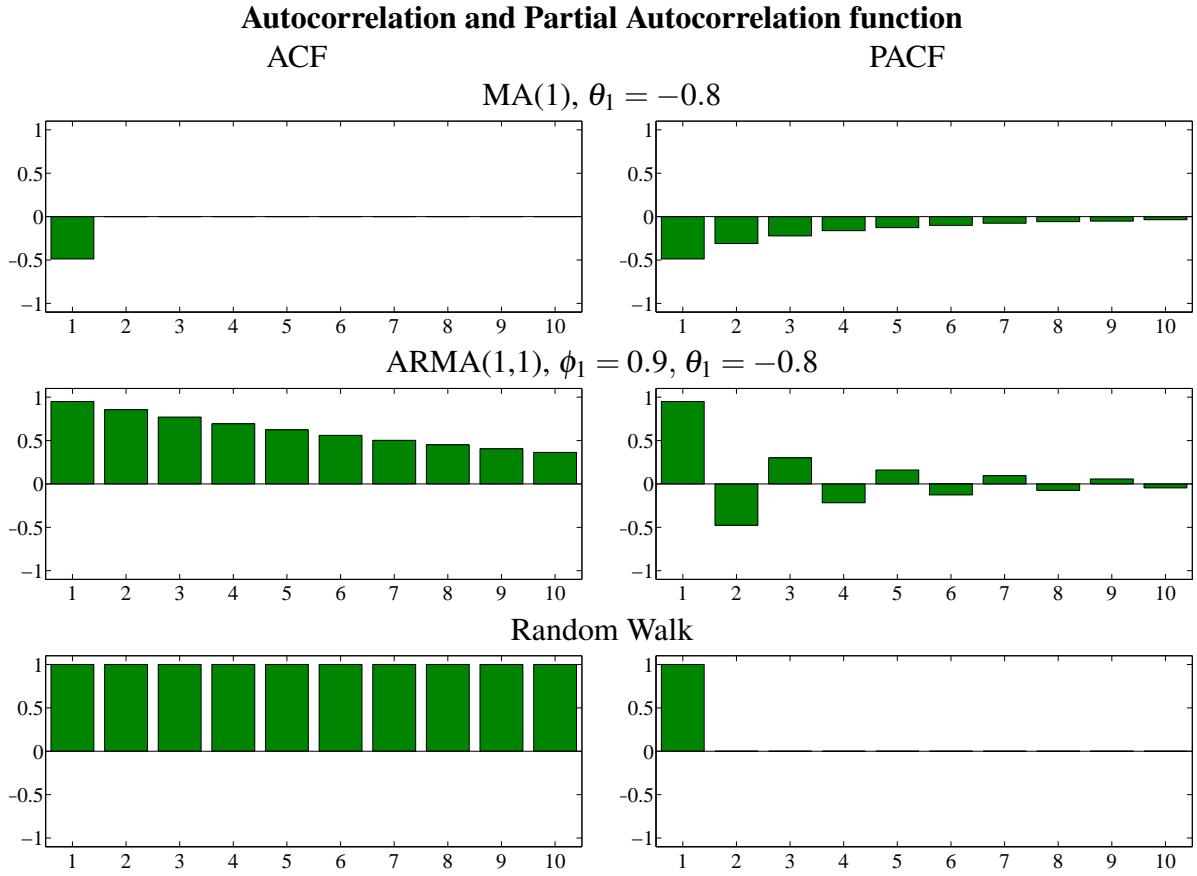


Figure 4.5: Autocorrelation function and partial autocorrelation function for 3 processes, an MA(1), and ARMA(1,1) and a random walk. Note the difference between how the ACF and PACF respond in AR and MA models.

$$\text{V}[\hat{\rho}_s] = T^{-1} \left(1 + 2 \sum_{j=1, j \neq s}^{\infty} \hat{\rho}_j^2 \right) \quad (4.55)$$

which is infeasible. The usual practice is to truncate the variance estimator at some finite lag L where L is a function of the sample size, often assumed that $L \propto T^{\frac{1}{3}}$ (if L is not an integer, rounding to the nearest one).⁸

Once the assumption of homoskedasticity is relaxed inference becomes more complicated. First consider the most restrictive null $H_0 : \rho_s = 0, s \neq 0$. If $\{y_t\}$ is a heteroskedastic white noise process (plus possibly a non-zero mean), inference can be made using White's heteroskedasticity robust covariance estimator (see chapter 3) so that

⁸The choice of $L \propto T^{\frac{1}{3}}$ is motivated by asymptotic theory where $T^{\frac{1}{3}}$ has been shown to be the optimal rate in the sense that it minimizes the asymptotic mean square error of the variance estimator.

$$\begin{aligned} V[\hat{\rho}_s] &= T^{-1} \left(T^{-1} \sum_{t=1}^T y_{t-s}^{*2} \right)^{-1} \left(T^{-1} \sum_{t=1}^T y_t^{*2} y_{t-s}^{*2} \right) \left(T^{-1} \sum_{t=1}^T y_{t-s}^{*2} \right)^{-1} \\ &= \frac{\sum_{t=s+1}^T y_t^{*2} y_{t-s}^{*2}}{\left(\sum_{t=s+1}^T y_{t-s}^{*2} \right)^2}. \end{aligned} \quad (4.56)$$

This covariance estimator is identical to White's covariance estimator for the regression

$$y_t = \rho_s y_{t-s} + \varepsilon_t$$

since under the null that $\rho_s = 0$, $y_t = \varepsilon_t$.

To test one of the more complicated null hypotheses a Heteroskedasticity-Autocorrelation Consistent (HAC) covariance estimator is required, the most common of which is the Newey-West covariance estimator.

Definition 4.21 (Newey-West Variance Estimator). Let z_t be a series that may be autocorrelated and define $z_t^* = z_t - \bar{z}$ where $\bar{z} = T^{-1} \sum_{t=1}^T z_t$. The L -lag Newey-West variance estimator for the variance of \bar{z} is

$$\begin{aligned} \hat{\sigma}_{NW}^2 &= T^{-1} \sum_{t=1}^T z_t^{*2} + 2 \sum_{l=1}^L w_l T^{-1} \sum_{t=l+1}^T z_t^* z_{t-l}^* \\ &= \hat{\gamma}_0 + 2 \sum_{l=1}^L w_l \hat{\gamma}_l \end{aligned} \quad (4.57)$$

where $\hat{\gamma}_l = T^{-1} \sum_{t=l+1}^T z_t^* z_{t-l}^*$ and $w_l = \frac{l+1-l}{L+1}$.

The Newey-West estimator has two important properties. First, it is always greater than 0. This is a desirable property of any variance estimator. Second, as long as $L \rightarrow \infty$, the $\hat{\sigma}_{NW}^2 \xrightarrow{P} V[y_t]$. The only remaining choice is which value to choose for L . Unfortunately this is problem dependent and it is important to use as small a value for L as the data will permit. Newey-West estimators tend to perform poorly in small samples and are worse, often substantially, than simpler estimators such as White's heteroskedasticity-consistent covariance estimator. This said, they also work in situations where White's estimator fails: when a sequence is autocorrelated White's estimator is not consistent.⁹ Long-run variance estimators are covered in more detail in the Multivariate Time Series chapter (chapter 5).

When used in a regression, the Newey-West estimator extends White's covariance estimator to allow $\{y_{t-s} \varepsilon_t\}$ to be both heteroskedastic and autocorrelated, setting $z_t^* = y_t^* y_{t-s}^*$,

⁹The Newey-West estimator nests White's covariance estimator as a special case by choosing $L = 0$.

$$\begin{aligned}
V[\hat{\rho}_s] &= T^{-1} \left(T^{-1} \sum_{t=s+1}^T y_{t-s}^{*2} \right)^{-1} \\
&\quad \times \left(T^{-1} \sum_{t=s+1}^T y_t^{*2} y_{t-s}^{*2} + 2 \sum_{j=1}^L w_j T^{-1} \sum_{t=s+j+1}^T y_t^* y_{t-s}^* (y_{t-j}^* y_{t-s-j}^*) \right) \\
&\quad \times \left(T^{-1} \sum_{t=s+1}^T y_{t-s}^{*2} \right)^{-1} \\
&= \frac{\sum_{t=s+1}^T y_t^{*2} y_{t-s}^{*2} + 2 \sum_{j=1}^L w_j \sum_{t=s+j+1}^T y_t^* y_{t-s}^* (y_{t-j}^* y_{t-s-j}^*)}{\left(\sum_{t=s+1}^T y_{t-s}^{*2} \right)^2}.
\end{aligned} \tag{4.58}$$

Note that only the center term has been changed and that L must diverge for this estimator to be consistent – even if $\{y_t\}$ follows an MA process, and the efficient choice sets $L \propto T^{\frac{1}{3}}$.

Tests that multiple autocorrelations are simultaneously zero can also be conducted. The standard method to test that s autocorrelations are zero, $H_0 = \rho_1 = \rho_2 = \dots = \rho_s = 0$, is the Ljung-Box Q statistic.

Definition 4.22 (Ljung-Box Q statistic). The Ljung-Box Q statistic, or simply Q statistic, tests the null that the first s autocorrelations are all zero against an alternative that at least one is non-zero: $H_0 : \rho_k = 0$ for $k = 1, 2, \dots, s$ versus $H_1 : \rho_k \neq 0$ for $k = 1, 2, \dots, s$. The test statistic is defined

$$Q = T(T+2) \sum_{k=1}^s \frac{\hat{\rho}_k^2}{T-k} \tag{4.59}$$

and Q has a standard χ_s^2 distribution.

The Q statistic is only valid under an assumption of homoskedasticity so caution is warranted when using it with financial data. A heteroskedasticity robust version of the Q -stat can be formed using an LM test.

Definition 4.23 (LM test for serial correlation). Under the null, $E[y_t^* y_{t-j}^*] = 0$ for $1 \leq j \leq s$. The LM-test for serial correlation is constructed by defining the score vector $\mathbf{s}_t = y_t^* [y_{t-1}^* y_{t-2}^* \dots y_{t-s}^*]'$,

$$LM = T \bar{\mathbf{s}}' \hat{\mathbf{S}} \bar{\mathbf{s}} \xrightarrow{d} \chi_s^2 \tag{4.60}$$

where $\bar{\mathbf{s}} = T^{-1} \sum_{t=1}^T \mathbf{s}_t$ and $\hat{\mathbf{S}} = T^{-1} \sum_{t=1}^T \mathbf{s}_t \mathbf{s}_t'$.¹⁰

Like the Ljung-Box Q statistic, this test has an asymptotic χ_s^2 distribution with the added advantage of being heteroskedasticity robust.

Partial autocorrelations can be estimated using regressions,

¹⁰Refer to chapters 2 and 3 for more on LM-tests.

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \hat{\phi}_s y_{t-s} + \varepsilon_t$$

where $\hat{\phi}_s = \hat{\phi}_s$. To test whether a partial autocorrelation is zero, the variance of $\hat{\phi}_s$, under the null and assuming homoskedasticity, is approximately T^{-1} for any s , and so a standard t -test can be used,

$$T^{\frac{1}{2}} \hat{\phi}_s \xrightarrow{d} N(0, 1). \quad (4.61)$$

If homoskedasticity cannot be assumed, White's covariance estimator can be used to control for heteroskedasticity.

Definition 4.24 (Sample Partial Autocorrelogram). A plot of the sample partial autocorrelations against the lag index known as a sample partial autocorrelogram.

4.6.3.1 Example: Autocorrelation, partial autocorrelation and Q Statistic

Figure 4.6 contains plots of the first 20 autocorrelations and partial autocorrelations of the VWM market returns and the default spread. The market appears to have a small amount of persistence and appears to be more consistent with a moving average than an autoregression. The default spread is highly persistent and appears to be a good candidate for an AR(1) since the autocorrelations decay slowly and the partial autocorrelations drop off dramatically after one lag, although an ARMA(1,1) cannot be ruled out.

4.6.4 Model Selection: The Box-Jenkins Methodology

The Box and Jenkins methodology is the most common approach for time-series model selection. It consists of two stages:

- Identification: Visual inspection of the series, the autocorrelations and the partial autocorrelations.
- Estimation: By relating the sample autocorrelations and partial autocorrelations to the ACF and PACF of ARMA models, candidate models are identified. These candidates are estimated and the residuals are tested for neglected dynamics using the residual autocorrelations, partial autocorrelations and Q statistics or LM-tests for serial correlation. If dynamics are detected in the residuals, a new model is specified and the procedure is repeated.

The Box-Jenkins procedure relies on two principles: parsimony and invertibility.

Definition 4.25 (Parsimony). Parsimony is a property of a model where the specification with the fewest parameters capable of capturing the dynamics of a time series is preferred to other representations equally capable of capturing the same dynamics.

Parsimony is an intuitive principle and using the smallest model has other benefits, particularly when forecasting. One consequence of the parsimony principle is that parameters which are not needed are excluded. For example, if the data generating process were an AR(1), selecting an AR(2) would adequately describe the process. The parsimony principle indicates the AR(1) should be preferred to an AR(2) since both are equally capable of capturing the dynamics of the data. Further, recall that an AR(1) can be reformulated as an MA(T) where $\theta_s = \phi_1^s$. Both the AR(1) and MA(T) are capable of

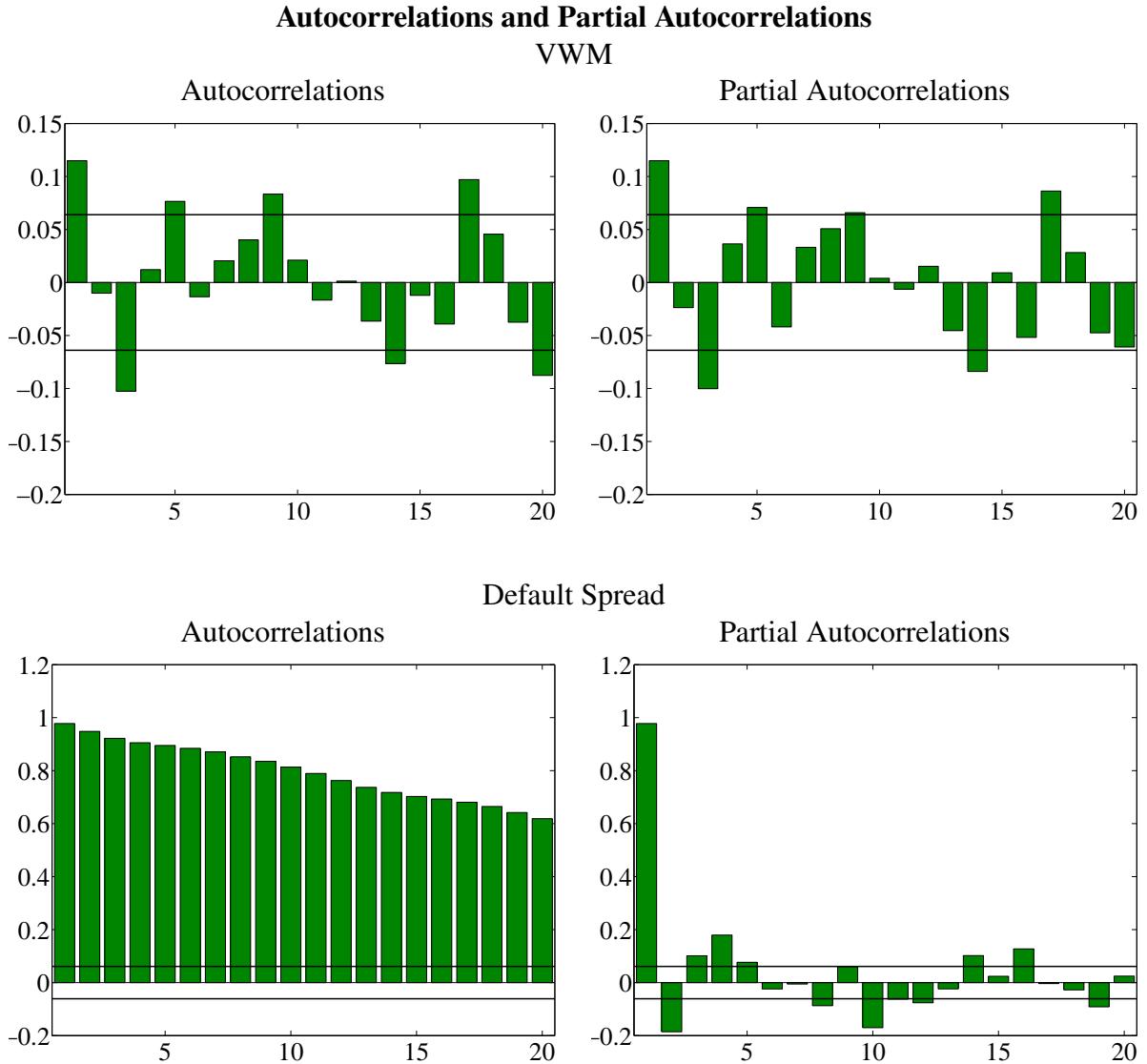


Figure 4.6: These four pictures plot the first 20 autocorrelations (left) and partial autocorrelations (right) of the VWM (top) and the Baa-Aaa spread (bottom). Approximate standard errors, assuming homoskedasticity, are in parenthesis.

capturing the dynamics of the data if the DGP is an AR(1), although the number of parameters in each is very different. The parsimony principle provides guidance on selecting the AR(1) over the MA(T) since it contains (many) fewer parameters yet provides an equivalent description of the relationship between current and past values of the data.

Definition 4.26 (Invertibility). A moving average is invertible if it can be written as a finite or convergent autoregression. Invertibility requires the roots of

$$(1 - \theta_1 z - \theta_2 z^2 - \dots - \theta_Q z^Q) = 0$$

to be greater than one in modulus (absolute value).

Invertibility is a technical requirement stemming from the use of the autocorrelogram and partial autocorrelogram to choose the model, and it plays an important role in achieving unique identification of the MA component of a model. For example, the ACF and PACF of

$$y_t = 2\epsilon_{t-1} + \epsilon_t$$

and

$$y_t = .5\epsilon_{t-1} + \epsilon_t$$

are identical. The first autocorrelation is $\theta_1/(1+\theta_1^2)$, and so in the first specification $\rho_1 = 2/(1+2^2) = .4$ and in the second $\rho_1 = .5/(1+.5^2) = .4$ while all other autocorrelations are zero. The partial autocorrelations are similarly identical – partial correlation are functions of autocorrelations – and so two processes are indistinguishable. Invertibility rules out the first of these two models since the root of $1 - 2z = 0$ is $\frac{1}{2} < 1$.

Information criteria such as the AIC or S/BIC can also be used to choose a model. Recall the definitions of the AIC and BIC:

Definition 4.27 (Akaike Information Criterion). The Akaike Information Criteria (AIC) is

$$AIC = \ln \hat{\sigma}^2 + \frac{2k}{T} \quad (4.62)$$

where $\hat{\sigma}^2$ is the estimated variance of the regression error and k is the number of parameters in the model.

Definition 4.28 (Schwarz/Bayesian Information Criterion). The Schwarz Information Criteria (SIC), also known as the Bayesian Information Criterion (BIC) is

$$BIC = \ln \hat{\sigma}^2 + \frac{k \ln T}{T} \quad (4.63)$$

where $\hat{\sigma}^2$ is the estimated variance of the regression error and k is the number of parameters in the model.

ICs are often applied by estimating the largest model which is thought to correctly capture the dynamics and then dropping lags until the AIC or S/BIC fail to decrease. Specific-to-General (StG) and General-to-Specific (GtS) are also applicable to time-series modeling and suffer from the same issues as those described in chapter 3, section 3.13.

4.7 Estimation

ARMA models are typically estimated using maximum likelihood (ML) estimation assuming that the errors are normal, using either conditional maximum likelihood, where the likelihood of y_t given y_{t-1}, y_{t-2}, \dots is used, or exact maximum likelihood where the joint distribution of $[y_1, y_2, \dots, y_{t-1}, y_t]$ is used.

4.7.1 Conditional Maximum Likelihood

Conditional maximum likelihood uses the distribution of y_t given y_{t-1}, y_{t-2}, \dots to estimate the parameters of an ARMA. The data are assumed to be conditionally normal, and so the likelihood is

$$\begin{aligned} f(y_t | y_{t-1}, y_{t-2}, \dots; \phi, \theta, \sigma^2) &= (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{\varepsilon_t^2}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{(y_t - \phi_0 - \sum_{i=1}^P \phi_i y_{t-i} - \sum_{j=1}^Q \theta_j \varepsilon_{t-j})^2}{2\sigma^2}\right) \end{aligned} \quad (4.64)$$

Since the $\{\varepsilon_t\}$ series is assumed to be a white noise process, the joint likelihood is simply the product of the individual likelihoods,

$$f(\mathbf{y}_t | \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \dots; \phi, \theta, \sigma^2) = \prod_{t=1}^T (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{\varepsilon_t^2}{2\sigma^2}\right) \quad (4.65)$$

and the conditional log-likelihood is

$$l(\phi, \theta, \sigma^2; \mathbf{y}_t | \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \dots) = -\frac{1}{2} \sum_{t=1}^T \ln 2\pi + \ln \sigma^2 + \frac{\varepsilon_t^2}{\sigma^2}. \quad (4.66)$$

Recall that the first-order condition for the mean parameters from a normal log-likelihood does not depend on σ^2 and that given the parameters in the mean equation, the maximum likelihood estimate of the variance is

$$\hat{\sigma}^2 = T^{-1} \sum_{t=1}^T (y_t - \phi_0 - \phi_1 y_{t-1} - \dots - \phi_P y_{t-P} - \theta_1 \varepsilon_{t-1} - \dots - \theta_Q \varepsilon_{t-Q})^2 \quad (4.67)$$

$$= T^{-1} \sum_{t=1}^T \varepsilon_t^2. \quad (4.68)$$

This allows the variance to be concentrated out of the log-likelihood so that it becomes

$$\begin{aligned} l(\mathbf{y}_t | \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \dots; \phi, \theta, \sigma^2) &= -\frac{1}{2} \sum_{t=1}^T \ln 2\pi + \ln(T^{-1} \sum_{t=1}^T \varepsilon_t^2) + \frac{\varepsilon_t^2}{T^{-1} \sum_{t=1}^T \varepsilon_t^2} \\ &= -\frac{1}{2} \sum_{t=1}^T \ln 2\pi - \frac{1}{2} \sum_{t=1}^T \ln(T^{-1} \sum_{t=1}^T \varepsilon_t^2) - \frac{T}{2} \sum_{t=1}^T \frac{\varepsilon_t^2}{\sum_{t=1}^T \varepsilon_t^2} \\ &= -\frac{1}{2} \sum_{t=1}^T \ln 2\pi - \frac{1}{2} \sum_{t=1}^T \ln(T^{-1} \sum_{t=1}^T \varepsilon_t^2) - \frac{T \sum_{t=1}^T \varepsilon_t^2}{2 \sum_{t=1}^T \varepsilon_t^2} \end{aligned} \quad (4.69)$$

$$\begin{aligned}
&= -\frac{1}{2} \sum_{t=1}^T \ln 2\pi - \frac{1}{2} \sum_{t=1}^T \ln(T^{-1} \sum_{t=1}^T \epsilon_t^2) - \frac{T}{2} \\
&= -\frac{1}{2} \sum_{t=1}^T \ln 2\pi - \frac{T}{2} - \frac{1}{2} \sum_{t=1}^T \ln(T^{-1} \sum_{t=1}^T \epsilon_t^2) \\
&= -\frac{1}{2} \sum_{t=1}^T \ln 2\pi - \frac{T}{2} - \frac{T}{2} \ln \hat{\sigma}^2.
\end{aligned}$$

Eliminating terms that do not depend on model parameters shows that maximizing the likelihood is equivalent to minimizing the error variance,

$$\max_{\phi, \theta, \sigma^2} l(\mathbf{y}_t | \mathbf{y}_{t-1}, \mathbf{y}_{t-2} \dots; \phi, \theta, \sigma^2) = -\frac{T}{2} \ln \hat{\sigma}^2. \quad (4.70)$$

where $\hat{\epsilon}_t = y_t - \phi_0 - \phi_1 y_{t-1} - \dots - \phi_P y_{t-P} - \theta_1 \epsilon_{t-1} - \dots - \theta_Q \epsilon_{t-Q}$, and so . estimation using conditional maximum likelihood is equivalent to least squares, although unlike linear regression the objective is nonlinear due to the moving average terms and so a nonlinear maximization algorithm is required. If the model does not include moving average terms ($Q = 0$), then the conditional maximum likelihood estimates of an AR(P) are identical the least squares estimates from the regression

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_P y_{t-P} + \epsilon_t. \quad (4.71)$$

Conditional maximum likelihood estimation of ARMA models requires either backcast values or truncation since some of the observations have low indices (e.g., y_1) that depend on observations not in the sample (e.g., $y_0, y_{-1}, \epsilon_0, \epsilon_{-1}$, etc.). Truncation is the most common and the likelihood is only computed for $t = P+1, \dots, T$, and initial values of ϵ_t are set to 0. When using backcasts, missing values of y can be initialized at the long-run average, $\bar{y} = T^{-1} \sum_{t=1}^T y_t$, and the initial values of ϵ_t are set to their unconditional expectation, 0. Using unconditional values works well when data are not overly persistent and T is not too small. The likelihood can then be recursively computed where estimated errors $\hat{\epsilon}_t$ used are using in moving average terms,

$$\hat{\epsilon}_t = y_t - \phi_0 - \phi_1 y_{t-1} - \dots - \phi_P y_{t-P} - \theta_1 \hat{\epsilon}_{t-1} - \dots - \theta_Q \hat{\epsilon}_{t-Q}, \quad (4.72)$$

where backcast values are used if any index is less than or equal to 0. The estimated residuals are then plugged into the conditional log-likelihood (eq. (4.69)) and the log-likelihood value is computed. The numerical maximizer will search for values of ϕ and θ that produce the largest log-likelihood. Once the likelihood optimizing values have been found, the maximum likelihood estimate of the variance is computed using

$$\hat{\sigma}^2 = T^{-1} \sum_{t=1}^T (y_t - \hat{\phi}_0 - \hat{\phi}_1 y_{t-1} - \dots - \hat{\phi}_P y_{t-P} - \hat{\theta}_1 \hat{\epsilon}_{t-1} - \dots - \hat{\theta}_Q \hat{\epsilon}_{t-Q})^2 \quad (4.73)$$

or the truncated version which sums from $P+1$ to T .

4.7.2 Exact Maximum Likelihood

Exact maximum likelihood directly utilizes the autocorrelation function of an ARMA(P,Q) to compute the correlation matrix of *all* of the y data, which allows the joint likelihood to be evaluated. Define

$$\mathbf{y} = [y_t \ y_{t-1} \ y_{t-2} \ \dots \ y_2 \ y_1]'$$

and let Γ be the T by T covariance matrix of \mathbf{y} . The joint likelihood of \mathbf{y} is given by

$$f(\mathbf{y}|\phi, \theta, \sigma^2) = (2\pi)^{-\frac{T}{2}} |\Gamma|^{-\frac{T}{2}} \exp\left(-\frac{\mathbf{y}'\Gamma^{-1}\mathbf{y}}{2}\right). \quad (4.74)$$

The log-likelihood is

$$l(\phi, \theta, \sigma^2; \mathbf{y}) = -\frac{T}{2} \ln(2\pi) - \frac{T}{2} \ln|\Gamma| - \frac{1}{2} \mathbf{y}'\Gamma^{-1}\mathbf{y}. \quad (4.75)$$

where Γ is a matrix of autocovariances,

$$\Gamma = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \gamma_3 & \dots & \gamma_{T-1} & \gamma_T \\ \gamma_1 & \gamma_0 & \gamma_1 & \gamma_2 & \dots & \gamma_{T-2} & \gamma_{T-1} \\ \gamma_2 & \gamma_1 & \gamma_0 & \gamma_1 & \dots & \gamma_{T-3} & \gamma_{T-2} \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ \gamma_{T-1} & \gamma_{T-2} & \gamma_{T-3} & \gamma_{T-4} & \dots & \gamma_0 & \gamma_1 \\ \gamma_T & \gamma_{T-1} & \gamma_{T-2} & \gamma_{T-3} & \dots & \gamma_1 & \gamma_0 \end{bmatrix}$$

and are determined by the model parameters (excluding the constant), ϕ , θ , and σ^2 . A nonlinear maximization algorithm can be used to search for the vector of parameters that maximizes this log-likelihood. The exact maximum likelihood estimator is generally believed to be more precise than conditional maximum likelihood and does not require backcasts of data or errors.

4.8 Inference

Inference on ARMA parameters from stationary time series is a standard application of maximum likelihood theory. Define $\psi = [\phi \ \theta \ \sigma^2]'$ as the parameter vector. Recall from 2 that maximum likelihood estimates are asymptotically normal,

$$\sqrt{T}(\psi - \hat{\psi}) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}^{-1}) \quad (4.76)$$

where

$$\mathcal{I} = -E\left[\frac{\partial^2 l(\mathbf{y}; \psi)}{\partial \psi \partial \psi'}\right].$$

where $\partial^2 l(\mathbf{y}; \psi)/\partial \psi \partial \psi'$ is the second derivative matrix of the log-likelihood (or Hessian). In practice \mathcal{I} is not known and it must be replaced with a consistent estimate,

$$\hat{\mathcal{I}} = T^{-1} \sum_{t=1}^T -\frac{\partial^2 l(y_t; \hat{\psi})}{\partial \psi \partial \psi'}.$$

Wald and t -tests on the parameter estimates can be computed using the elements of \mathcal{I} , or likelihood ratio tests can be used by imposing the null on the model and comparing the log-likelihood values of the constrained and unconstrained estimators.

One important assumption in the above distribution theory is that the estimator is a maximum likelihood estimator; this requires the likelihood to be correctly specified, or, in other words, for the data to be homoskedastic and normally distributed. This is generally an implausible assumption when using financial data and a modification of the above theory is needed. When one likelihood is specified for the data but they actually have a different distribution the estimator is known as a Quasi Maximum Likelihood estimator (QML). QML estimators, like ML estimators, are asymptotically normal under mild regularity conditions on the data but with a different asymptotic covariance matrix,

$$\sqrt{T}(\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}^{-1} \mathcal{J} \mathcal{I}^{-1}) \quad (4.77)$$

where

$$\mathcal{J} = E \left[\frac{\partial l(\mathbf{y}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \frac{\partial l(\mathbf{y}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}'} \right]$$

\mathcal{J} must also be estimated and the usual estimator is

$$\hat{\mathcal{J}} = T^{-1} \sum_{t=1}^T \frac{\partial l(y_t; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \frac{\partial l(y_t; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}'}$$

where $\frac{\partial l(y_t; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}}$ is the score of the log-likelihood. $\mathcal{I}^{-1} \mathcal{J} \mathcal{I}^{-1}$ is known as a sandwich covariance estimator, White's covariance estimator.

A sandwich covariance estimator is needed when the model for the data is not completely specified or is misspecified, and it accounts for the failure of Information Matrix Inequality to hold (see chapters 2 and 3). As was the case in linear regression, a sufficient condition for the IME to fail in ARMA estimation is heteroskedastic residuals. Considering the prevalence of conditionally heteroskedasticity in financial data, this is nearly a given.

4.9 Forecasting

Forecasting is a common objective of many time-series models. The objective of a forecast is to minimize a loss function.

Definition 4.29 (Loss Function). A loss function is a function of the observed data, y_{t+h} and the time- t constructed forecast, $\hat{y}_{t+h|t}$, $L(y_{t+h}, \hat{y}_{t+h|t})$, that has the three following properties:

- Property 1: The loss of any forecast is non-negative, so $L(y_{t+h}, \hat{y}_{t+h|t}) \geq 0$.
- Property 2: There exists a point, y_{t+h}^* , known as the optimal forecast, where the loss function takes the value 0. That is $L(y_{t+h}, y_{t+h}^*) = 0$.
- Property 3: The loss is non-decreasing away from y_{t+h}^* . That is if $y_{t+h}^B > y_{t+h}^A > y_{t+h}^*$, then $L(y_{t+h}, y_{t+h}^B) > L(y_{t+h}, y_{t+h}^A) > L(y_{t+h}, y_{t+h}^*)$. Similarly, if $y_{t+h}^D < y_{t+h}^C < y_{t+h}^*$, then $L(y_{t+h}, y_{t+h}^D) > L(y_{t+h}, y_{t+h}^C) > L(y_{t+h}, y_{t+h}^*)$.

The most common loss function is Mean Square Error (MSE) which chooses the forecast to minimize

$$\text{E}[L(y_{t+h}, \hat{y}_{t+h|t})] = \text{E}[(y_{t+h} - \hat{y}_{t+h|t})^2] \quad (4.78)$$

where $\hat{y}_{t+h|t}$ is the time- t forecast of y_{t+h} . Notice that this is just the optimal projection problem and the optimal forecast is the conditional mean, $y_{t+h|t}^* = \text{E}_t[y_{t+h}]$ (See chapter 3). It is simple to verify that this loss function satisfies the properties of a loss function. Property 1 holds by inspection and property 2 occurs when $y_{t+h} = \hat{y}_{t+h|t}^*$. Property 3 follows from the quadratic form. MSE is far and away the most common loss function but others, such as Mean Absolute Deviation (MAD), Quad-Quad and Linex are used both in practice and in the academic literature. The MAD loss function will be revisited in chapter 6 (Value-at-Risk). The Advanced Financial Econometrics elective will study non-MSE loss functions in more detail.

The remainder of this section will focus exclusively on forecasts that minimize the MSE loss function. Fortunately, in this case, forecasting from ARMA models is an easy exercise. For simplicity consider the AR(1) process,

$$y_t = \phi_0 + \phi_1 y_{t-1} + \varepsilon_t.$$

Since the optimal forecast is the conditional mean, all that is needed is to compute $\text{E}_t[y_{t+h}]$ for any h . When $h = 1$,

$$y_{t+1} = \phi_0 + \phi_1 y_t + \varepsilon_{t+1}$$

so the conditional expectation is

$$\begin{aligned} \text{E}_t[y_{t+1}] &= \text{E}_t[\phi_0 + \phi_1 y_t + \varepsilon_{t+1}] \\ &= \phi_0 + \phi_1 \text{E}_t[y_t] + \text{E}_t[\varepsilon_{t+1}] \\ &= \phi_0 + \phi_1 y_t + 0 \\ &= \phi_0 + \phi_1 y_t \end{aligned} \quad (4.79)$$

which follows since y_t is in the time- t information set (\mathcal{F}_t) and $\text{E}_t[\varepsilon_{t+1}] = 0$ by assumption.¹¹ The optimal forecast for $h = 2$ is given by $\text{E}_t[y_{t+2}]$,

$$\begin{aligned} \text{E}_t[y_{t+2}] &= \text{E}_t[\phi_0 + \phi_1 y_{t+1} + \varepsilon_{t+2}] \\ &= \phi_0 + \phi_1 \text{E}_t[y_{t+1}] + \text{E}_t[\varepsilon_{t+2}] \\ &= \phi_0 + \phi_1 (\phi_0 + \phi_1 y_t) + 0 \\ &= \phi_0 + \phi_1 \phi_0 + \phi_1^2 y_t \end{aligned}$$

which follows by substituting in the expression derived in eq. (4.79) for $\text{E}_t[y_{t+1}]$. The optimal forecast for any arbitrary h uses the recursion

$$\text{E}_t[y_{t+h}] = \phi_0 + \phi_1 \text{E}_t[y_{t+h-1}] \quad (4.80)$$

¹¹This requires a slightly stronger assumption than ε_t is a white noise process.

and it is easily shown that $E_t[y_{t+h}] = \phi_0 \sum_{i=0}^{h-1} \phi_1^i + \phi_1^h y_t$. If $|\phi_1| < 1$, as $h \rightarrow \infty$, the forecast of y_{t+h} and $E_t[y_{t+h}]$ converges to $\phi_0/(1 - \phi_1)$, the unconditional expectation of y_t . In other words, for forecasts in the distant future there is no information about the location of y_{t+h} other than it will return to its unconditional mean. This is not surprising since y_t is covariance stationary when $|\phi_1| < 1$.

Next consider forecasts from an MA(2),

$$y_t = \phi_0 + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \varepsilon_t.$$

The one-step-ahead forecast is given by

$$\begin{aligned} E_t[y_{t+1}] &= E_t[\phi_0 + \theta_1 \varepsilon_t + \theta_2 \varepsilon_{t-1} + \varepsilon_{t+1}] \\ &= \phi_0 + \theta_1 E_t[\varepsilon_t] + \theta_2 E_t[\varepsilon_{t-1}] + E_t[\varepsilon_{t+1}] \\ &= \phi_0 + \theta_1 \varepsilon_t + \theta_2 \varepsilon_{t-1} + 0 \end{aligned}$$

which follows since ε_t and ε_{t-1} are in the \mathcal{F}_t information set and $E_t[\varepsilon_{t+1}] = 0$ by assumption. In practice the one step ahead forecast would be given by

$$E_t[y_{t+1}] = \hat{\phi}_0 + \hat{\theta}_1 \hat{\varepsilon}_t + \hat{\theta}_2 \hat{\varepsilon}_{t-1}$$

where both the unknown parameters *and* the unknown residuals would be replaced with their estimates.¹² The 2-step ahead forecast is given by

$$\begin{aligned} E_t[y_{t+2}] &= E_t[\phi_0 + \theta_1 \varepsilon_{t+1} + \theta_2 \varepsilon_t + \varepsilon_{t+2}] \\ &= \phi_0 + \theta_1 E_t[\varepsilon_{t+1}] + \theta_2 E_t[\varepsilon_t] + E_t[\varepsilon_{t+2}] \\ &= \phi_0 + \theta_1 0 + \theta_2 \varepsilon_t + 0 \\ &= \phi_0 + \theta_2 \varepsilon_t. \end{aligned}$$

The 3 or higher step forecast can be easily seen to be ϕ_0 . Since all future residuals have zero expectation they cannot affect long horizon forecasts. Like the AR(1) forecast, the MA(2) forecast is mean reverting. Recall the unconditional expectation of an MA(Q) process is simply ϕ_0 . For any $h > Q$ the forecast of y_{t+h} is just this value, ϕ_0 .

Finally consider the 1 to 3-step ahead forecasts from an ARMA(2,2),

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \varepsilon_t.$$

Conditioning on the information set \mathcal{F}_t , the expectation of y_{t+1} is

$$\begin{aligned} E_t[y_{t+1}] &= E_t[\phi_0 + \phi_1 y_t + \phi_2 y_{t-1} + \theta_1 \varepsilon_t + \theta_2 \varepsilon_{t-1} + \varepsilon_{t+1}] \\ &= E_t[\phi_0] + E_t[\phi_1 y_t] + E_t[\phi_2 y_{t-1}] + E_t[\theta_1 \varepsilon_t] + E_t[\theta_2 \varepsilon_{t-1}] + E_t[\varepsilon_{t+1}]. \end{aligned}$$

Noting that all of the elements are in \mathcal{F}_t except ε_{t+1} , which has conditional expectation 0,

¹²The residuals are a natural by-product of the parameter estimation stage.

$$E_t[y_{t+1}] = \phi_0 + \phi_1 y_t + \phi_2 y_{t-1} + \theta_1 \varepsilon_t + \theta_2 \varepsilon_{t-1}$$

Note that in practice, the parameters and errors will all be replaced by their estimates (i.e. $\hat{\phi}_1$ and $\hat{\varepsilon}_t$). The 2-step ahead forecast is given by

$$\begin{aligned} E_t[y_{t+2}] &= E_t[\phi_0 + \phi_1 y_{t+1} + \phi_2 y_t + \theta_1 \varepsilon_{t+1} + \theta_2 \varepsilon_t + \varepsilon_{t+2}] \\ &= E_t[\phi_0] + E_t[\phi_1 y_{t+1}] + E_t[\phi_2 y_t] + \theta_1 E_t[\varepsilon_{t+1}] + \theta_2 E_t[\varepsilon_t] + E_t[\varepsilon_{t+2}] \\ &= \phi_0 + \phi_1 E_t[y_{t+1}] + \phi_2 y_t + \theta_1 E_t[\varepsilon_{t+1}] + \theta_2 \varepsilon_t + E_t[\varepsilon_{t+2}] \\ &= \phi_0 + \phi_1 (\phi_0 + \phi_1 y_t + \phi_2 y_{t-1} + \theta_1 \varepsilon_t + \theta_2 \varepsilon_{t-1}) + \phi_2 y_t + \theta_1 0 + \theta_2 \varepsilon_t + 0 \\ &= \phi_0 + \phi_1 \phi_0 + \phi_1^2 y_t + \phi_1 \phi_2 y_{t-1} + \phi_1 \theta_1 \varepsilon_t + \phi_1 \theta_2 \varepsilon_{t-1} + \phi_2 y_t + \theta_2 \varepsilon_t \\ &= \phi_0 + \phi_1 \phi_0 + (\phi_1^2 + \phi_2) y_t + \phi_1 \phi_2 y_{t-1} + (\phi_1 \theta_1 + \theta_2) \varepsilon_t + \phi_1 \theta_2 \varepsilon_{t-1}. \end{aligned}$$

In this case, there are three terms which are not known at time t . By assumption $E_t[\varepsilon_{t+2}] = E_t[\varepsilon_{t+1}] = 0$ and $E_t[y_{t+1}]$ has been computed above, so

$$E_t[y_{t+2}] = \phi_0 + \phi_1 \phi_0 + (\phi_1^2 + \phi_2) y_t + \phi_1 \phi_2 y_{t-1} + (\phi_1 \theta_1 + \theta_2) \varepsilon_t + \phi_1 \theta_2 \varepsilon_{t-1}$$

In a similar manner,

$$\begin{aligned} E_t[y_{t+3}] &= \phi_0 + \phi_1 E_t[y_{t+2}] + \phi_2 E_t[y_{t+1}] + \theta_1 \varepsilon_{t+2} + \theta_2 \varepsilon_{t+1} + \varepsilon_{t+3} \\ E_t[y_{t+3}] &= \phi_0 + \phi_1 E_t[y_{t+2}] + \phi_2 E_t[y_{t+1}] + 0 + 0 + 0 \end{aligned}$$

which is easily solved by plugging in the previously computed values for $E_t[y_{t+2}]$ and $E_t[y_{t+1}]$. This pattern can be continued by iterating forward to produce the forecast for an arbitrary h .

Two things are worth noting from this discussion:

- If there is no AR component, all forecast for $h > Q$ will be ϕ_0 .
- For large h , the optimal forecast converges to the unconditional expectation given by

$$\lim_{h \rightarrow \infty} E_t[y_{t+h}] = \frac{\phi_0}{1 - \phi_1 - \phi_2 - \dots - \phi_P} \quad (4.81)$$

4.9.1 Forecast Evaluation

Forecast evaluation is an extensive topic and these notes only cover two simple yet important tests: Mincer-Zarnowitz regressions and Diebold-Mariano tests.

4.9.1.1 Mincer-Zarnowitz Regressions

Mincer-Zarnowitz regressions (henceforth MZ) are used to test for the optimality of the forecast and are implemented with a standard regression. If a forecast is correct, it should be the case that a regression of the realized value on its forecast and a constant should produce coefficients of 1 and 0 respectively.

Definition 4.30 (Mincer-Zarnowitz Regression). A Mincer-Zarnowitz (MZ) regression is a regression of a forecast, $\hat{y}_{t+h|t}$ on the realized value of the predicted variable, y_{t+h} and a constant,

$$y_{t+h} = \beta_1 + \beta_2 \hat{y}_{t+h|t} + \eta_t. \quad (4.82)$$

If the forecast is optimal, the coefficients in the MZ regression should be consistent with $\beta_1 = 0$ and $\beta_2 = 1$.

For example, let $\hat{y}_{t+h|t}$ be the h -step ahead forecast of y constructed at time t . Then running the regression

$$y_{t+h} = \beta_1 + \beta_2 \hat{y}_{t+h|t} + v_t$$

should produce estimates close to 0 and 1. Testing is straightforward and can be done with any standard test (Wald, LR or LM). An augmented MZ regression can be constructed by adding time- t measurable variables to the original MZ regression.

Definition 4.31 (Augmented Mincer-Zarnowitz Regression). An Augmented Mincer-Zarnowitz regression is a regression of a forecast, $\hat{y}_{t+h|t}$ on the realized value of the predicted variable, y_{t+h} , a constant and any other time- t measurable variables, $\mathbf{x}_t = [x_{1t} x_{2t} \dots x_{Kt}]$,

$$y_{t+h} = \beta_1 + \beta_2 \hat{y}_{t+h|t} + \beta_3 x_{1t} + \dots + \beta_{K+2} x_{Kt} + \eta_t. \quad (4.83)$$

If the forecast is optimal, the coefficients in the MZ regression should be consistent with $\beta_1 = \beta_3 = \dots = \beta_{K+2} = 0$ and $\beta_2 = 1$.

It is crucial that the additional variables are time- t measurable and are in \mathcal{F}_t . Again, any standard test statistic can be used to test the null $H_0 : \beta_2 = 1 \cap \beta_1 = \beta_3 = \dots = \beta_{K+2} = 0$ against the alternative $H_1 : \beta_2 \neq 1 \cup \beta_j \neq 0, j = 1, 3, 4, \dots, K-1, K-2$.

4.9.1.2 Diebold-Mariano Tests

A Diebold-Mariano test, in contrast to an MZ regression, examines the relative performance of two forecasts. Under MSE, the loss function is given by $L(y_{t+h}, \hat{y}_{t+h|t}) = (y_{t+h} - \hat{y}_{t+h|t})^2$. Let A and B index the forecasts from two models $\hat{y}_{t+h|t}^A$ and $\hat{y}_{t+h|t}^B$, respectively. The losses from each can be defined as $l_t^A = (y_{t+h} - \hat{y}_{t+h|t}^A)^2$ and $l_t^B = (y_{t+h} - \hat{y}_{t+h|t}^B)^2$. If the models were equally good (or bad), one would expect $\bar{l}^A \approx \bar{l}^B$ where \bar{l} is the average loss. If model A is better, meaning it has a lower expected loss $E[L(y_{t+h}, \hat{y}_{t+h|t}^A)] < E[L(y_{t+h}, \hat{y}_{t+h|t}^B)]$, then, on average, it should be the case that $\bar{l}^A < \bar{l}^B$. Alternatively, if model B were better it should be the case that $\bar{l}^B < \bar{l}^A$. The DM test exploits this to construct a simple t -test of equal predictive ability.

Definition 4.32 (Diebold-Mariano Test). Define $d_t = l_t^A - l_t^B$. The Diebold-Mariano test is a test of equal predictive accuracy and is constructed as

$$DM = \frac{\bar{d}}{\sqrt{\widehat{V}[\bar{d}]}}$$

where M (for modeling) is the number of observations used in the model building and estimation, R (for reserve) is the number of observations held back for model evaluation and $\bar{d} = R^{-1} \sum_{t=M+1}^{M+R} d_t$. Under the null that $E[L(y_{t+h}, \hat{y}_{t+h|t}^A)] = E[L(y_{t+h}, \hat{y}_{t+h|t}^B)]$, and under some regularity conditions on $\{d_t\}$, $DM \xrightarrow{d} N(0, 1)$. $V[d_t]$ is the *long-run variance* of d_t and must be computed using a HAC covariance estimator.

If the models are equally accurate, one would expect that $E[d_t] = 0$ which forms the null of the DM test, $H_0 : E[d_t] = 0$. To test the null, a standard t -stat is used although the test has two alternatives: $H_1^A : E[d_t] < 0$ and $H_1^B : E[d_t] > 0$ which correspond to the superiority of model A or B , respectively. DM is asymptotically normally distributed. Large negative values (less than -2) indicate model A produces less loss on average and hence is superior, while large positive values indicate the opposite. Values close to zero indicate neither is statistically superior.

In Diebold-Marino tests the variance must be estimated using a Heteroskedasticity-Autocorrelation Consistent variance estimator.

Definition 4.33 (Heteroskedasticity Autocorrelation Consistent Covariance Estimator). Covariance estimators which are robust to both ignored autocorrelation in residuals and to heteroskedasticity are known as Heteroskedasticity-Autocorrelation Consistent (HAC) covariance. The most common example of an HAC estimator is the Newey-West (or Bartlett) covariance estimator.

The typical variance estimator cannot be used in DM tests and a kernel estimator must be substituted (e.g., Newey-West).

Despite all of these complications, implementing a DM test is very easy. The first step is to compute the series of losses, $\{l_t^A\}$ and $\{l_t^B\}$, for both forecasts. Next compute $d_t = l_t^A - l_t^B$. Finally, regress d_t on a constant and use Newey-West errors,

$$d_t = \beta_1 + \varepsilon_t.$$

The t -stat on β_1 is the DM test statistic and can be compared to critical values of a normal distribution.

4.10 Nonstationary Time Series

Nonstationary time series present some particular difficulties and standard inference often fails when a process depends explicitly on t . Nonstationarities can be classified into one of four categories:

- Seasonalities
- Deterministic Trends (also known as Time Trends)
- Unit Roots (also known as Stochastic Trends)
- Structural Breaks

Each type has a unique feature. Seasonalities are technically a form of deterministic trend, although their analysis is sufficiently similar to stationary time series that little is lost in treating a seasonal time series as if it were stationary. Processes with deterministic trends have unconditional means which depend on time while unit roots processes have unconditional variances that grow over time. Structural breaks are an encompassing class which may result in either or both the mean and variance exhibiting time dependence.

4.10.1 Seasonality, Diurnality, and Hebdomadality

Seasonality, diurnality and hebdomadality are pervasive in economic time series. While many data series have been *seasonally adjusted* to remove seasonalities, particularly US macroeconomic series, there are many time-series where no seasonally adjusted version is available. Ignoring seasonalities is detrimental to the precision of parameters and forecasting and model estimation and selection is often more precise when both seasonal and nonseasonal dynamics are simultaneously modeled.

Definition 4.34 (Seasonality). Data are said to be seasonal if they exhibit a non-constant deterministic pattern with an annual frequency.

Definition 4.35 (Hebdomadality). Data which exhibit day-of-week deterministic effects are said to be hebdomadal.

Definition 4.36 (Diurnality). Data which exhibit intra-daily deterministic effects are said to be diurnal.

Seasonal data are non-stationary, although seasonally de-trended data (usually referred to as de-seasonalized data) may be stationary. Seasonality is common in macroeconomic time series, diurnality is pervasive in ultra-high frequency data (tick data) and hebdomadality is often believed to be a feature of asset prices. Seasonality is, technically, a form of non-stationarity since the mean of a process exhibits explicit dependence on t through the seasonal component, and the Box-Jenkins methodology is not directly applicable. However, a slight change in time scale, where the seasonal pattern is directly modeled along with any non-seasonal dynamics produces a residual series which is stationary and so the Box-Jenkins methodology may be applied.

For example, consider a seasonal quarterly time series. Seasonal dynamics may occur at lags $4, 8, 12, 16, \dots$, while nonseasonal dynamics can occur at any lag $1, 2, 3, 4, \dots$. Note that multiples of 4 appear in both lists and so the identification of the seasonal and nonseasonal dynamics may be difficult (although separate identification makes little practical difference).

The standard practice when working with seasonal data is to conduct model selection over two sets of lags by choosing a maximum lag to capture the seasonal dynamics and by choosing a maximum lag to capture nonseasonal ones. Returning to the example of a seasonal quarterly time series, a model may need to examine up to 4 lags to capture nonseasonal dynamics and up to 4 lags of the seasonal component, and if the seasonal component is annual, these four seasonal lags correspond to regressors as $t - 4, t - 8, t - 12$, and $t - 16$.

4.10.1.1 Example: Seasonality

Most U.S. data series are available seasonally adjusted, something that is not true for data from many areas of the world, including the Euro zone. This example makes use of monthly data on the U.S. money supply, M1, a measure of the money supply that includes all coins, currency held by the public, travelers' checks, checking account balances, NOW accounts, automatic transfer service accounts, and balances in credit unions.

Figure 4.10.1.1 contains a plot of monthly M1, the growth of M1 (log differences), and the sample autocorrelogram and sample partial autocorrelogram of M1. These figures show evidence of an annual seasonality (lags 12, 24 and 36), and applying the Box-Jenkins methodology, the seasonality appears to be a seasonal AR, or possibly a seasonal ARMA. The short run dynamics oscillate and

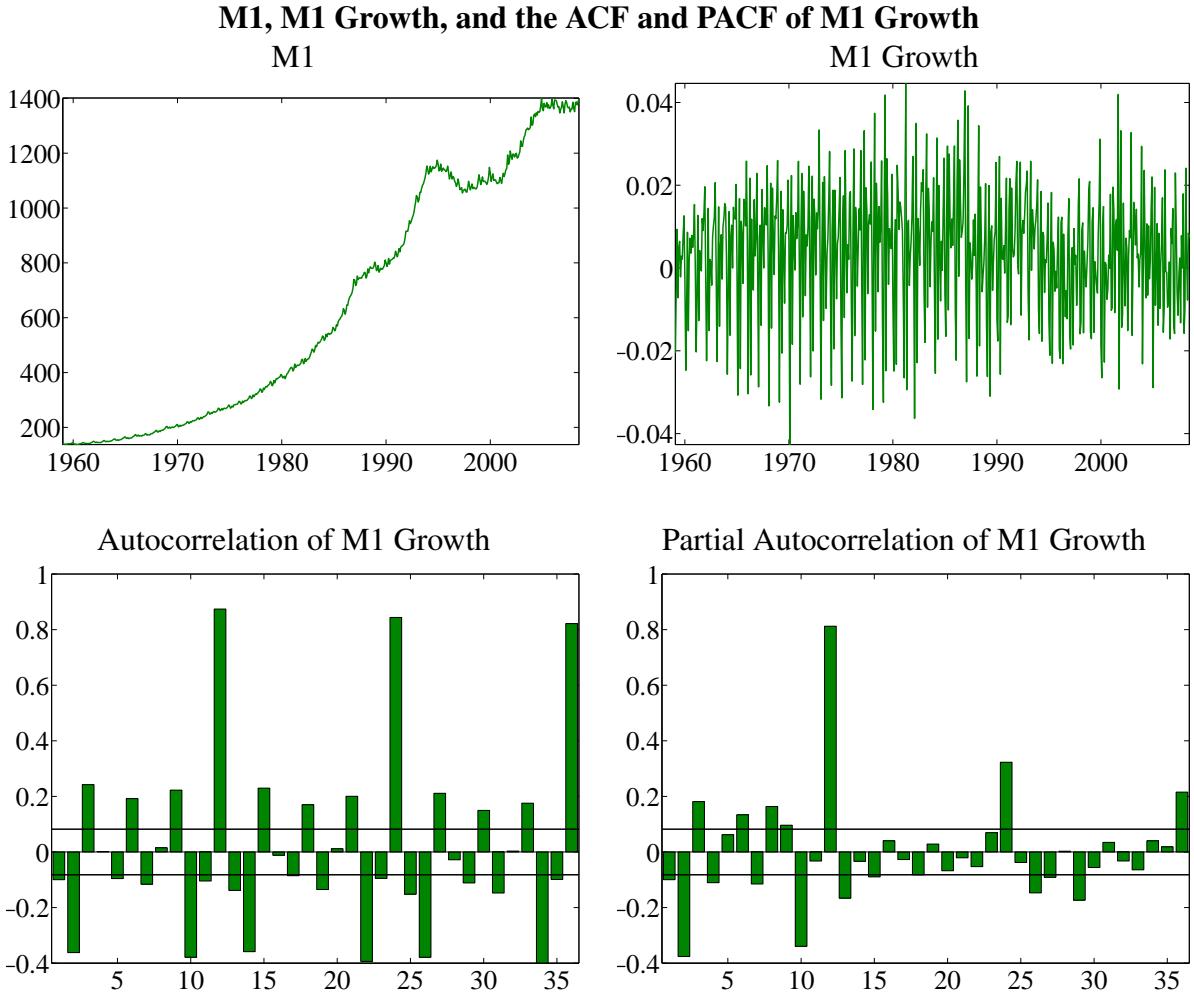


Figure 4.7: Plot of the money supply (M1), M1 growth (log differences), and the sample autocorrelogram and sample partial autocorrelogram of M1 growth. There is a clear seasonal pattern at 12 months which appears consistent with a seasonal ARMA(1,1).

appear consistent with an autoregression since the partial autocorrelations are fairly flat (aside from the seasonal component). Three specifications which may be appropriate to model the process were fit: a 12 month seasonal AR, a 12 month seasonal MA and a 12-month seasonal ARMA, all combined with an AR(1) to model the short run dynamics. Results are reported in table 4.3

4.10.2 Deterministic Trends

The simplest form of nonstationarity is a deterministic trend. Models with deterministic time trends can be decomposed into three components:

$$y_t = \text{deterministic trend} + \text{stationary component} + \text{noise} \quad (4.84)$$

where $\{y_t\}$ would be stationary if the trend were absent. The two most common forms of time trends are polynomial (linear, quadratic, etc) and exponential. Processes with polynomial time trends can be

Modeling seasonalities in M1 growth

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_{12} y_{t-12} + \theta_{12} \varepsilon_{t-12} + \varepsilon_t$$

$\hat{\phi}_0$	$\hat{\phi}_1$	$\hat{\phi}_{12}$	$\hat{\theta}_{12}$	SIC
0.000 (0.245)	-0.014 (0.000)	0.984 (0.000)	-0.640 (0.000)	-9.989
0.001 (0.059)	-0.011 (0.000)	0.873 (0.000)		-9.792
0.004 (0.002)	-0.008 (0.000)		0.653 (0.000)	-9.008

Table 4.3: Estimated parameters, p-values and SIC for three models with seasonalities. The SIC prefers the larger specification with both seasonal AR and MA terms. Moreover, correctly modeling the seasonalities frees the AR(1) term to model the oscillating short run dynamics (notice the significant negative coefficient).

expressed

$$y_t = \phi_0 + \delta_1 t + \delta_2 t^2 + \dots + \delta_S t^S + \text{stationary component} + \text{noise},$$

and linear time trend models are the most common,

$$y_t = \phi_0 + \delta_1 t + \text{stationary component} + \text{noise}.$$

For example, consider a linear time trend model with an MA(1) stationary component,

$$y_t = \phi_0 + \delta_1 t + \theta_1 \varepsilon_{t-1} + \varepsilon_t$$

The long-run behavior of this process is dominated by the time trend, although it may still exhibit persistent fluctuations around $\delta_1 t$.

Exponential trends appear as linear or polynomial trends in the log of the dependent variable, for example

$$\ln y_t = \phi_0 + \delta_1 t + \text{stationary component} + \text{noise}.$$

The trend is the permanent component of a nonstationary time series, and so any two observations are permanently affected by the trend line irrespective of the number of observations between them. The class of deterministic trend models can be reduced to a stationary process by detrending.

Definition 4.37 (Trend Stationary). A stochastic process, $\{y_t\}$ is trend stationary if there exists a nontrivial function $g(t, \delta)$ such that $\{y_t - g(t, \delta)\}$ is stationary.

Detrended data may be strictly or covariance stationary (or both).

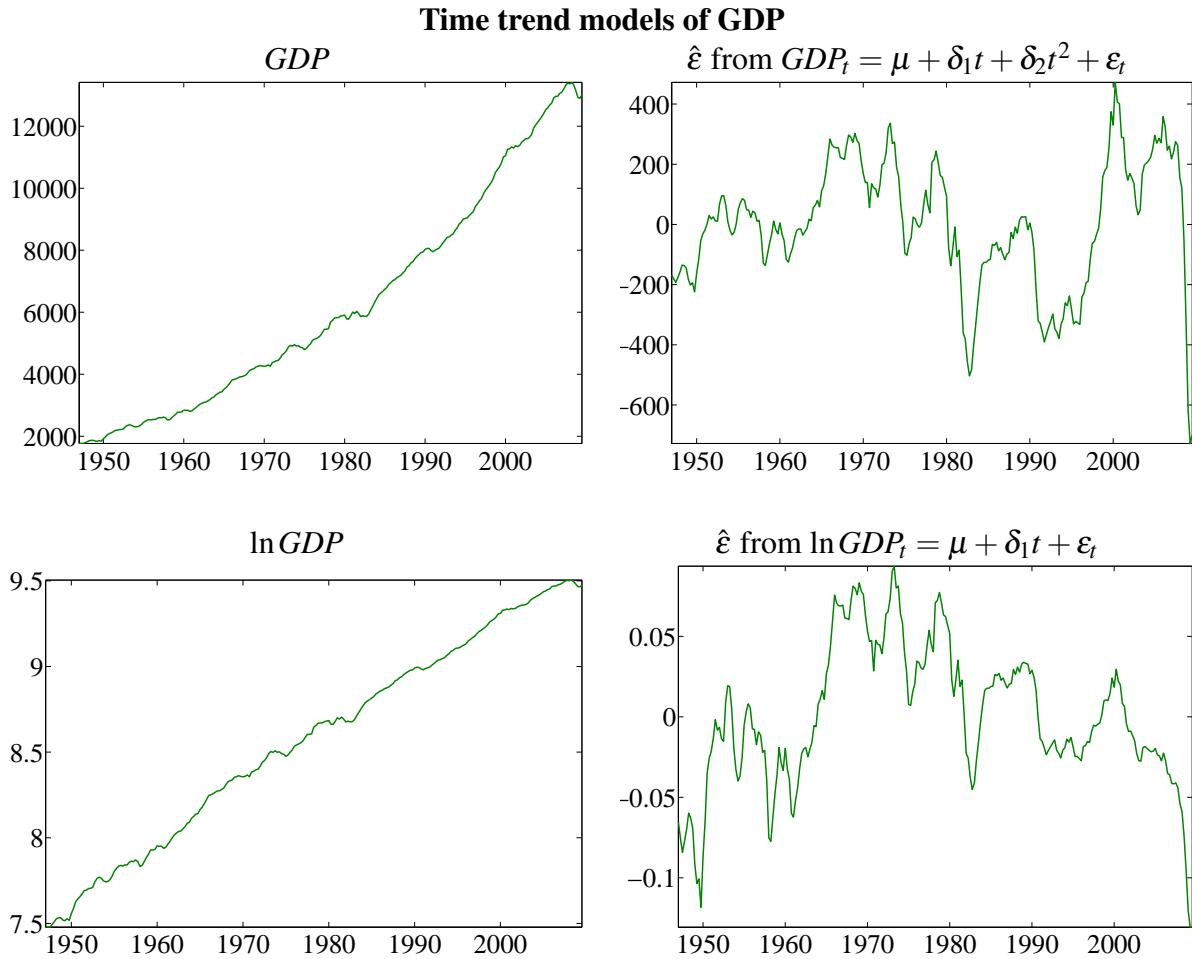


Figure 4.8: Two time trend models are presented, one on the levels of GDP and one on the natural log. Note that the detrended residuals are still highly persistent. This is a likely sign of a unit root.

4.10.2.1 Modeling the time trend in GDP

U.S. GDP data was taken from FRED II from Q1 1947 until Q2 July 2008. To illustrate the use of a time trend, consider two simple models for the level of GDP. The first models the level as a quadratic function of time while the second models the natural log of GDP in an exponential trend model.

$$GDP_t = \phi_0 + \delta_1 t + \delta_2 t^2 + \varepsilon_t$$

and

$$\ln GDP_t = \phi_0 + \delta_1 t + \varepsilon_t.$$

Figure 4.8 presents the time series of GDP, the log of GDP and errors from two models that include trends. Neither time trend appears to remove the extreme persistence in GDP which may indicate the process contains a unit root.

4.10.3 Unit Roots

Unit root processes are generalizations of the classic random walk. A process is said to have a unit root if the distributed lag polynomial can be factored so that one of the roots is exactly one.

Definition 4.38 (Unit Root). A stochastic process, $\{y_t\}$, is said to contain a unit root if

$$(1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_P L^P) y_t = \phi_0 + (1 - \theta_1 L - \theta_2 L^2 - \dots - \theta_Q L^Q) \varepsilon_t \quad (4.85)$$

can be factored

$$(1 - L)(1 - \tilde{\phi}_1 L - \tilde{\phi}_2 L^2 - \dots - \tilde{\phi}_{P-1} L^{P-1}) y_t = \phi_0 + (1 - \theta_1 L - \theta_2 L^2 - \dots - \theta_Q L^Q) \varepsilon_t. \quad (4.86)$$

The simplest example of a unit root process is a random walk.

Definition 4.39 (Random Walk). A stochastic process $\{y_t\}$ is known as a random walk if

$$y_t = y_{t-1} + \varepsilon_t \quad (4.87)$$

where ε_t is a white noise process with the additional property that $E_{t-1}[\varepsilon_t] = 0$.

The basic properties of a random walk are simple to derive. First, a random walk is a martingale since $E_t[y_{t+h}] = y_t$ for any h .¹³ The variance of a random walk can be deduced from

$$\begin{aligned} V[y_t] &= E[(y_t - y_0)^2] \\ &= E[(\varepsilon_t + y_{t-1} - y_0)^2] \\ &= E[(\varepsilon_t + \varepsilon_{t-1} + y_{t-2} - y_0)^2] \\ &= E[(\varepsilon_t + \varepsilon_{t-1} + \dots + \varepsilon_1)^2] \\ &= E[\varepsilon_t^2 + \varepsilon_{t-1}^2 + \dots + \varepsilon_1^2] \\ &= t\sigma^2 \end{aligned} \quad (4.88)$$

and this relationship holds for any time index, and so $V[y_s] = s\sigma^2$. The s^{th} autocovariance (γ_s) of a unit root process is given by

$$\begin{aligned} V[(y_t - y_0)(y_{t-s} - y_0)] &= E[(\varepsilon_t + \varepsilon_{t-1} + \dots + \varepsilon_1)(\varepsilon_{t-s} + \varepsilon_{t-s-1} + \dots + \varepsilon_1)] \\ &= E[(\varepsilon_{t-s}^2 + \varepsilon_{t-s-1}^2 + \dots + \varepsilon_1^2)] \\ &= (t-s)\sigma^2 \end{aligned} \quad (4.89)$$

and the s^{th} autocorrelation is then

$$\rho_s = \frac{t-s}{t} \quad (4.90)$$

¹³Since the effect of an innovation never declines in a unit root process, it is not reasonable to consider the infinite past as in a stationary AR(1).

which tends to 1 for large t and fixed s . This is a useful property of a random walk process (and any unit root process): The autocorrelations will be virtually constant at 1 with only a small decline at large lags. Building from the simple unit root, one can define a unit root plus drift model,

$$y_t = \delta + y_{t-1} + \varepsilon_t$$

which can be equivalently expressed

$$y_t = \delta t + \sum_{i=1}^t \varepsilon_i + y_0$$

and so the random walk plus drift process consists of both a deterministic trend and a random walk. Alternatively, a random walk model can be augmented with stationary noise so that

$$y_t = \sum_{i=1}^t \varepsilon_i + \eta_t$$

which leads to the general class of random walk models plus stationary noise processes

$$\begin{aligned} y_t &= \sum_{i=1}^t \varepsilon_i + \sum_{j=1}^{t-1} \theta_j \eta_{t-j} + \eta_t \\ &= \sum_{i=1}^t \varepsilon_i + \Theta(L) \eta_t \end{aligned}$$

where $\Theta(L)\eta_t = \sum_{j=1}^{t-1} \theta_j \eta_{t-j} + \eta_t$ is a compact expression for a lag polynomial in θ . Since $\Theta(L)\eta_t$ can include any covariance stationary process, this class should be considered general. More importantly, this process has two components: a permanent one, $\sum_{i=1}^t \varepsilon_i$ and a transitory one $\Theta(L)\eta_t$. The permanent behaves similarly to a deterministic time trend, although unlike the deterministic trend model, the permanent component of this specification depends on random increments. For this reason, it is known as a *stochastic trend*.

Like the deterministic model, where the process can be detrended, a process with a unit root can be stochastically detrended, or *differenced*, $\Delta y_t = y_t - y_{t-1}$. Differencing a random walk produces a stationary series,

$$\begin{aligned} y_t - y_{t-1} &= \sum_{i=1}^t \varepsilon_i + \Theta(L) \eta_t - \sum_{i=1}^{t-1} \varepsilon_i + \Theta(L) \eta_{t-1} \\ \Delta y_t &= \varepsilon_t + (1-L)\Theta(L) \eta_t \end{aligned}$$

Over-differencing occurs when the difference operator is applied to a stationary series. While over-differencing cannot create a unit root, it does have negative consequences such as increasing the variance of the residual and reducing the magnitude of possibly important dynamics. Finally, unit root processes are often known as I(1) processes.

Definition 4.40 (Integrated Process of Order 1). A stochastic process $\{y_t\}$ is integrated of order 1, written $I(1)$, if $\{y_t\}$ is non-covariance-stationary and if $\{\Delta y_t\}$ is covariance stationary. Note: A process that is already covariance stationary is said to be $I(0)$.

The expression integrated is derived from the presence of $\sum_{i=1}^t \varepsilon_i$ in a unit root process where the sum operator is the discrete version of an integrator.

4.10.4 Difference or Detrend?

Detrending removes nonstationarities from deterministically trending series while differencing removes stochastic trends from unit roots. What happens if the wrong type of detrending is used? The unit root case is simple, and since the trend is stochastic, no amount of detrending can eliminate the permanent component. Only knowledge of the stochastic trend at an earlier point in time can transform the series to be stationary.

Differencing a stationary series produces another series which is stationary but with a larger variance than a detrended series.

$$\begin{aligned} y_t &= \delta t + \varepsilon_t \\ \Delta y_t &= \delta + \varepsilon_t - \varepsilon_{t-1} \end{aligned}$$

while the properly detrended series would be

$$y_t - \delta t = \varepsilon_t$$

If ε_t is a white noise process, the variance of the differenced series is twice that of the detrended series with a large negative MA component. The parsimony principle dictates that the correctly detrended series should be preferred even though differencing is a viable method of transforming a nonstationary series to be stationary. Higher orders of time trends can be eliminated by re-differencing at the cost of even higher variance.

4.10.5 Testing for Unit Roots: The Dickey-Fuller Test and the Augmented DF Test

Dickey-Fuller tests (DF), and their generalization to augmented Dickey-Fuller tests (ADF) are the standard test for unit roots. Consider the case of a simple random walk,

$$y_t = y_{t-1} + \varepsilon_t$$

so that

$$\Delta y_t = \varepsilon_t.$$

Dickey and Fuller noted that if the null of a unit root were true, then

$$y_t = \phi_1 y_{t-1} + \varepsilon_t$$

can be transformed into

$$\Delta y_t = \gamma y_{t-1} + \varepsilon_t$$

where $\gamma = \phi - 1$ and a test could be conducted for the null $H_0 : \gamma = 0$ against an alternative $H_1 : \gamma < 0$. This test is equivalent to testing whether $\phi = 1$ in the original model. $\hat{\gamma}$ can be estimated using a simple regression of Δy_t on y_{t-1} , and the t -stat can be computed in the usual way. If the distribution of $\hat{\gamma}$ were standard normal (under the null), this would be a very simple test. Unfortunately, it is non-standard since, under the null, y_{t-1} is a unit root and the variance is growing rapidly as the number of observations increases. The solution to this problem is to use the Dickey-Fuller distribution rather than the standard normal to make inference on the t -stat of $\hat{\gamma}$.

Dickey and Fuller considered three separate specifications for their test,

$$\begin{aligned}\Delta y_t &= \gamma y_{t-1} + \varepsilon_t \\ \Delta y_t &= \phi_0 + \gamma y_{t-1} + \varepsilon_t \\ \Delta y_t &= \phi_0 + \delta_1 t + \gamma y_{t-1} + \varepsilon_t\end{aligned}\tag{4.91}$$

which correspond to a unit root, a unit root with a linear time trend, and a unit root with a quadratic time trend. The null and alternative hypotheses are the same: $H_0 : \gamma = 0$, $H_1 : \gamma < 0$ (one-sided alternative), and the null that y_t contains a unit root will be rejected if $\hat{\gamma}$ is sufficiently negative, which is equivalent to $\hat{\phi}$ being significantly less than 1 in the original specification.

Unit root testing is further complicated since the inclusion of deterministic regressor(s) affects the asymptotic distribution. For example, if $T = 200$, the critical values of a Dickey-Fuller distribution are

	No trend	Linear	Quadratic
10%	-1.66	-2.56	-3.99
5%	-1.99	-2.87	-3.42
1%	-2.63	-3.49	-3.13

The Augmented Dickey-Fuller (ADF) test generalized the DF to allow for short-run dynamics in the differenced dependent variable. The ADF is a DF regression augmented with lags of the differenced dependent variable to capture short-term fluctuations around the stochastic trend,

$$\begin{aligned}\Delta y_t &= \gamma y_{t-1} + \sum_{p=1}^P \phi_p \Delta y_{t-p} + \varepsilon_t \\ \Delta y_t &= \phi_0 + \gamma y_{t-1} + \sum_{p=1}^P \phi_p \Delta y_{t-p} + \varepsilon_t \\ \Delta y_t &= \phi_0 + \delta_1 t + \gamma y_{t-1} + \sum_{p=1}^P \phi_p \Delta y_{t-p} + \varepsilon_t\end{aligned}\tag{4.92}$$

Neither the null and alternative hypotheses nor the critical values are changed by the inclusion of lagged dependent variables. The intuition behind this result stems from the observation that the Δy_{t-p} are “less integrated” than y_t and so are asymptotically less informative.

4.10.6 Higher Orders of Integration

In some situations, integrated variables are not just I(1) but have a higher order or integration. For example, the log consumer price index ($\ln CPI$) is often found to be I(2) (integrated of order 2) and so *double differencing* is required to transform the original data into a stationary series. As a consequence, both the level of $\ln CPI$ and its difference (inflation) contain unit roots.

Definition 4.41 (Integrated Process of Order d). A stochastic process $\{y_t\}$ is integrated of order d , written $I(d)$, if $\{(1 - L)^d y_t\}$ is a covariance stationary ARMA process.

Testing for higher orders of integration is simple: repeat the DF or ADF test on the differenced data. Suppose that it is not possible to reject the null that the level of a variable, y_t , is integrated and so the data should be differenced (Δy_t). If the differenced data rejects a unit root, the testing procedure is complete and the series is consistent with an I(1) process. If the differenced data contains evidence of a unit root, the data should be double differenced ($\Delta^2 y_t$) and the test repeated. The null of a unit root should be rejected on the double-differenced data since no economic data are thought to be I(3), and so if the null cannot be rejected on double-differenced data, careful checking for omitted deterministic trends or other serious problems in the data is warranted.

4.10.6.1 Power of Unit Root tests

Recall that the power of a test is 1 minus the probability Type II error, or simply the probability that the null is rejected when the null is false. In the case of a unit root, power is the ability of a test to reject the null that the process contains a unit root when the largest characteristic root is less than 1. Many economic time-series have roots close to 1 and so it is important to maximize the power of a unit root test so that models have the correct order of integration.

DF and ADF tests are known to be very sensitive to misspecification and, in particular, have very low power if the ADF specification is not flexible enough to account for factors other than the stochastic trend. Omitted deterministic time trends or insufficient lags of the differenced dependent variable both lower the power by increasing the variance of the residual. This works analogously to the classic regression testing problem of having a low power test when the residual variance is too large due to omitted variables.

A few recommendations can be made regarding unit root tests:

- Use a loose model selection criteria to choose the lag length of the included differenced dependent variables (e.g., AIC).
- Including extraneous deterministic regressors lowers power, but failing to include relevant deterministic regressors produces a test with no power, even asymptotically, and so be conservative when excluding deterministic regressors.
- More powerful tests than the ADF are available. Specifically, DF-GLS of Elliott, Rothenberg, and Stock (1996) is increasingly available and it has maximum power against certain alternatives.
- Trends tend to be obvious and so always plot both the data and the differenced data.

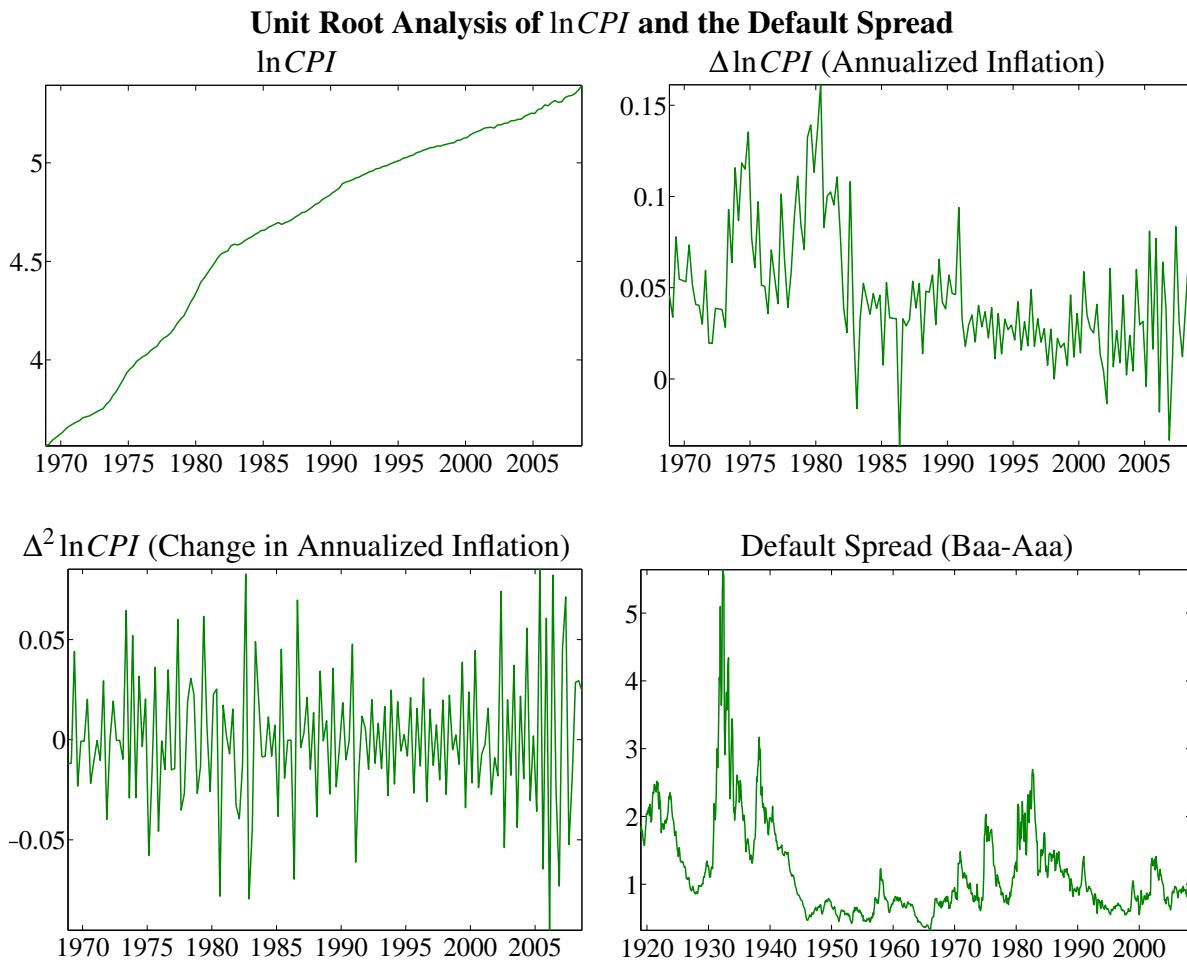


Figure 4.9: These four panels plot the log consumer price index ($\ln CPI$), $\Delta \ln CPI$, $\Delta^2 \ln CPI$ and the default spread. Both $\Delta^2 \ln CPI$ and the default spread reject the null of a unit root.

- Use a general-to-specific search to perform unit root testing. Start from a model which should be too large. If the unit root is rejected, one can be confident that there is not a unit root since this is a low power test. If a unit root cannot be rejected, reduce the model by removing insignificant deterministic components first since these lower power without affecting the t -stat. If all regressors are significant, and the null still cannot be rejected, then conclude that the data contains a unit root.

4.10.7 Example: Unit root testing

Two series will be examined for unit roots: the default spread and the log U.S. consumer price index. The $\ln CPI$, which measure consumer prices index less energy and food costs (also known as core inflation), has been taken from FRED, consists of quarterly data and covers the period between August 1968 and August 2008. Figure 4.9 contains plots of both series as well as the first and second differences of $\ln CPI$.

	$\ln CPI$	$\ln CPI$	$\ln CPI$	$\Delta \ln CPI$	$\Delta \ln CPI$	$\Delta^2 \ln CPI$	Default Sp.	Default Sp.
t -stat	-2.119	-1.541	1.491	-2.029	-0.977	-13.535	-3.130	-1.666
p-val	0.543	0.517	0.965	0.285	0.296	0.000	0.026	0.091
Deterministic	Linear	Const.	None	Const.	None	None	Const.	None
# lags	4	4	4	3	3	2	15	15

Table 4.4: ADF results for tests that $\ln CPI$ and the default spread have unit roots. The null of an unit root cannot be rejected in $\ln CPI$, nor can the null that $\Delta \ln CPI$ contains a unit root and so CPI appears to be an I(2) process. The default spread rejects the null of a unit root although clearly highly persistent.

$\ln CPI$ is trending and the spread does not have an obvious time trend. However, deterministic trends should be over-specified and so the initial model for $\ln CPI$ will include both a constant and a time-trend and the model for the spread will include a constant. The lag length used in the model was automatically selected using the BIC.

Results of the unit root tests are presented in table 4.4. Based on this output, the spreads reject a unit root at the 5% level but the $\ln CPI$ cannot. The next step is to difference the $\ln CPI$ to produce $\Delta \ln CPI$. Rerunning the ADF test on the differenced CPI (inflation) and including either a constant or no deterministic trend, the null of a unit root still cannot be rejected. Further differencing the data, $\Delta^2 \ln CPI_t = \delta \ln CPI_t - \ln CPI_{t-1}$, strongly rejects, and so $\ln CPI$ appears to be I(2). The final row of the table indicates the number of lags used in the ADF and was selected using the BIC with a maximum of 12 lags for $\ln CPI$ or 36 lags for the spread (3 years).

4.11 Nonlinear Models for Time-Series Analysis

While this chapter has exclusively focused on linear time-series processes, many non-linear time-series processes have been found to parsimoniously describe important dynamics in financial data. Two which have proven particularly useful in the analysis of financial data are Markov Switching Autoregressions (MSAR) and Threshold Autoregressions (TAR), especially the subclass of Self-Exciting Threshold Autoregressions (SETAR).¹⁴

4.12 Filters

The ultimate goal of most time-series modeling is to forecast a time-series in its entirety, which requires a model for both permanent and transitory components. In some situations, it may be desirable to focus on either the short-run dynamics or the long-run dynamics exclusively, for example in technical analysis where prices are believed to be long-run unpredictable but may have some short- or medium-run predictability. Linear filters are a class of functions which can be used to “extract” a stationary cyclic component from a time-series which contains both short-run dynamics and a permanent component. Generically, a filter for a time series $\{y_t\}$ is defined as

¹⁴There are many nonlinear models frequently used in financial econometrics for modeling quantities other than the conditional mean. For example, both the ARCH (conditional volatility) and CaViaR (conditional Value-at-Risk) models are nonlinear in the original data.

$$x_t = \sum_{i=-\infty}^{\infty} w_i y_{t-i} \quad (4.93)$$

where x_t is the filtered time-series or filter output. In most applications, it is desirable to assign a label to x_t , either a trend, written τ_t , or a cyclic component, c_t .

Filters are further categorized into *causal* and *non-causal*, where causal filters are restricted to depend on only the past and present of y_t , and so as a class are defined through

$$x_t = \sum_{i=0}^{\infty} w_i y_{t-i}. \quad (4.94)$$

Causal filters are more useful in isolating trends from cyclical behavior for forecasting purposes while non-causal filters are more useful for historical analysis of macroeconomic and financial data.

4.12.1 Frequency, High- and Low-Pass Filters

This text has exclusively dealt with time series in the *time domain* – that is, understanding dynamics and building models based on the time distance between points. An alternative strategy for describing a time series is in terms of *frequencies* and the magnitude of the cycle at a given frequency. For example, suppose a time series has a cycle that repeats every 4 periods. This series could be equivalently described as having a cycle that occurs with a frequency of 1 in 4, or .25. A frequency description is relatively compact – it is only necessary to describe a process from frequencies of 0 to 0.5, the latter of which would be a cycle with a period of 2.¹⁵

The idea behind filtering is to choose a set of frequencies and then to isolate the cycles which occur within the selected frequency range. Filters that eliminate high-frequency cycles are known as *low-pass filters*, while filters that eliminate low-frequency cycles are known as *high-pass filters*. Moreover, high- and low-pass filters are related in such a way that if $\{w_i\}$ is a set of weights corresponding to a high-pass filter, $v_0 = 1 - w_0$, $v_i = -w_i$ $i \neq 0$ is a set of weights corresponding to a low-pass filter. This relationship forms an identity since $\{v_i + w_i\}$ must correspond to an *all-pass* filter since $\sum_{i=-\infty}^{\infty} (v_i + w_i) y_{t-i} = y_t$ for any set of weights.

The goal of a filter is to select a particular frequency range and nothing else. The *gain function* describes the amount of attenuations which occurs at a given frequency.¹⁶ A gain of 1 at a particular frequency means any signal at that frequency is passed through unmodified while a gain of 0 means that the signal at that frequency is eliminated from the filtered data. Figure 4.10 contains a graphical representation of the gain function for a set of *ideal filters*. The four panels show an all-pass (all frequencies unmodified), a low-pass filter with a cutoff frequency of $\frac{1}{10}$, a high-pass with a cutoff

¹⁵The frequency $\frac{1}{2}$ is known as the *Nyquist* frequency since it is not possible to measure any cyclic behavior at frequencies above $\frac{1}{2}$ since these would have a cycle of 1 period and so would appear constant.

¹⁶The gain function for any filter of the form $x_t = \sum_{i=-\infty}^{\infty} w_i y_{t-i}$ can be computed as

$$G(f) = \left| \sum_{k=-\infty}^{\infty} w_k \exp(-ik2\pi f) \right|$$

where $i = \sqrt{-1}$.

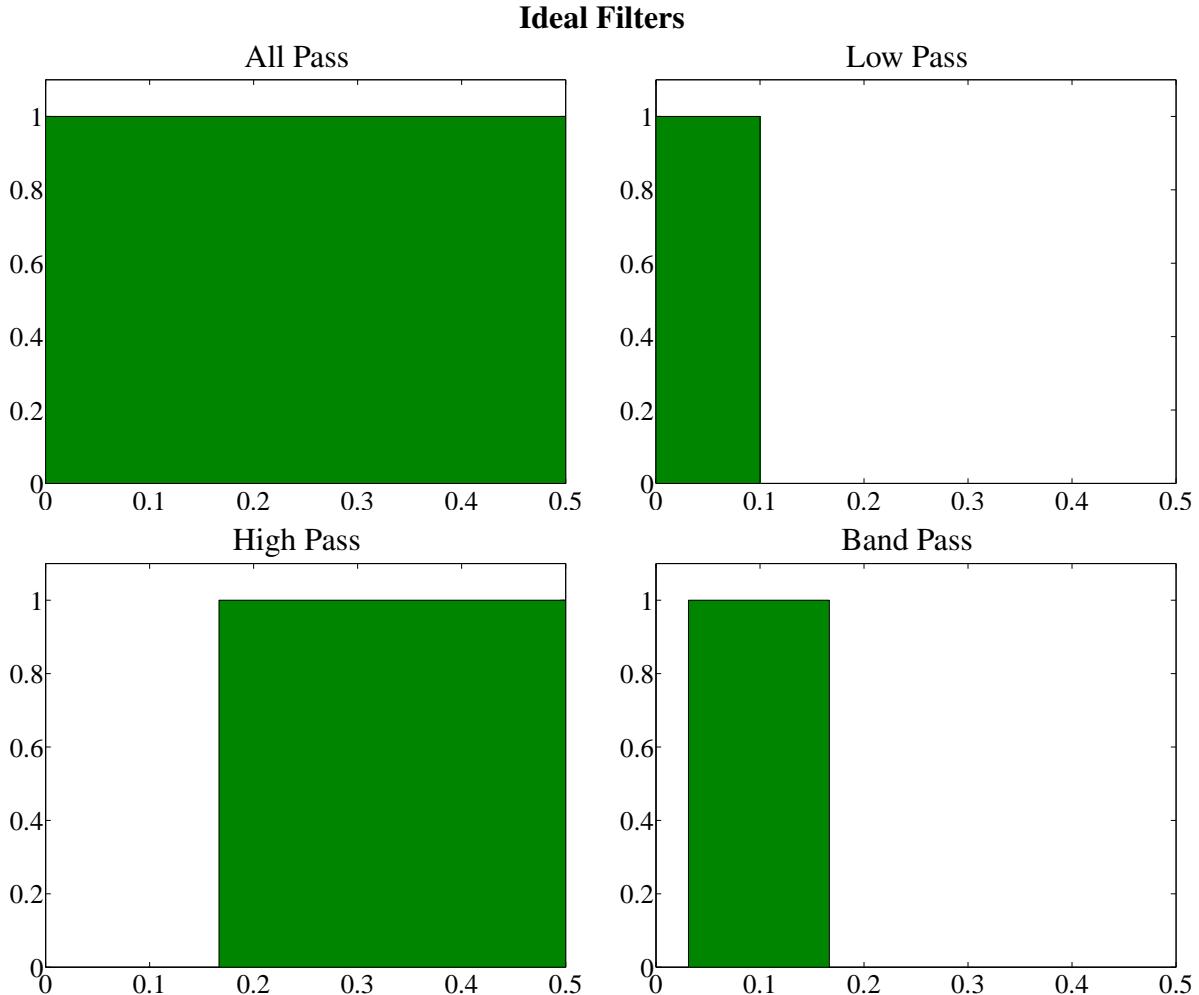


Figure 4.10: These four panels depict the gain functions from a set of *ideal filters*. The all-pass filter allows all frequencies through. The low-pass filter cuts off at $\frac{1}{10}$. The high-pass cuts off below $\frac{1}{6}$ and the band-pass filter cuts off below $\frac{1}{32}$ and above $\frac{1}{6}$.

frequency of $\frac{1}{6}$, and a band-pass filter with cutoff frequencies of $\frac{1}{6}$ and $\frac{1}{32}$.¹⁷ In practice, only the all-pass filter (which corresponds to a filter with weights $w_0 = 1, w_i = 0$ for $i \neq 0$) can be constructed using a finite sum, and so applied filtering must make trade-offs.

4.12.2 Moving Average and Exponentially Weighted Moving Average (EWMA)

Moving averages are the simplest filters and are often used in technical analysis. Moving averages can be constructed as both causal and non-causal filters.

Definition 4.42 (Causal Moving Average). A causal moving average (MA) is a function which takes

¹⁷Band-pass filters are simply the combination of two low-pass filters. Specifically, if $\{w_i\}$ is set of weights from a low-pass filter with a cutoff frequency of f_1 and $\{v_i\}$ is a set of weights from a low-pass filter with a cutoff frequency of f_2 , $f_2 > f_1$, then $\{v_i - w_i\}$ is a set of weights which corresponds to a band-pass filter with cutoffs at f_1 and f_2 .

the form

$$\tau_t = \frac{1}{n} \sum_{i=1}^n y_{t-i+1}.$$

Definition 4.43 (Centered (Non-Causal) Moving Average). A centered moving average (MA) is a function which takes the form

$$\tau_t = \frac{1}{2n+1} \sum_{i=-n}^n y_{t-i+1}.$$

Note that the centered MA is an average over $2n + 1$ data points.

Moving averages are low-pass filters since their weights add up to 1. In other words, the moving average would contain the permanent component of $\{y_t\}$ and so would have the same order of integration. The cyclic component, $c_t = y_t - \tau_t$, would have a lower order of integration than y_t . Since moving averages are low-pass filters, the difference of two moving averages must be a band-pass filter. Figure 4.11 contains the gain function from the difference between a 20-day and 50-day moving average which is commonly used in technical analysis.

Exponentially Weighted Moving Averages (EWMA) are a close cousin of the MA which places greater weight on recent data than on past data.

Definition 4.44 (Exponentially Weighted Moving Average). A exponentially weighed moving average (EWMA) is a function which takes the form

$$\tau_t = (1 - \lambda) \sum_{i=0}^{\infty} \lambda^i y_{t-i}$$

for some $\lambda \in (0, 1)$.

The name EWMA is derived from the exponential decay in the weights, and EWMA can be equivalently expressed (up to an initial condition) as

$$\tau_t = (1 - \lambda) \lambda y_t + \lambda \tau_{t-1}.$$

Like MAs, EWMA are low-pass filters since the weights sum to 1.

EWMA are commonly used in financial applications as volatility filters, where the dependent variable is chosen to be the squared return. The difference between two EWMA is often referred to as a Moving Average Convergence Divergence (MACD) filter in technical analysis. MACDs are indexed by two numbers, a fast period and a slow period, where the number of data in the MACD can be converted to λ as $\lambda = (n - 1)/(n + 1)$, and so an MACD(12,26) is the difference between two EWMA with parameters .842 and .926. 4.11 contains the gain function from an MACD(12,26) (the difference between two EWMA), which is similar to, albeit smoother than, the gain function from the difference of a 20-day and a 50-day MAs.

4.12.3 Hodrick-Prescott Filter

The Hodrick and Prescott (1997) (HP) filter is constructed by balancing the fitting the trend to the data and the smoothness of the trend. The HP filter is defined as the solution to

$$\min_{\tau_t} \sum_{t=1}^T (y_t - \tau_t)^2 + \lambda \sum_{t=2}^{T-1} ((\tau_{t-1} - \tau_t) - (\tau_t + \tau_{t-1}))$$

where $(\tau_{t-1} - \tau_t) - (\tau_t + \tau_{t-1})$ can be equivalently expressed as the second-difference of τ_t , $\Delta^2 \tau_t$. λ is a smoothness parameter: if $\lambda = 0$ then the solution to the problem is $\tau_t = y_t \forall t$, and as $\lambda \rightarrow \infty$, the “cost” of variation in $\{\tau_t\}$ becomes arbitrarily high and $\tau_t = \beta_0 + \beta_1 t$ where β_0 and β_1 are the least squares fit of a linear trend model to the data.

It is simple to show that the solution to this problem must have

$$\mathbf{y} = \Gamma \boldsymbol{\tau}$$

where Γ is a band-diagonal matrix (all omitted elements are 0) of the form

$$\Gamma = \begin{bmatrix} 1+\lambda & -2\lambda & \lambda & & & & \\ -2\lambda & 1+5\lambda & -4\lambda & \lambda & & & \\ \lambda & -4\lambda & 1+6\lambda & -4\lambda & \lambda & & \\ & \lambda & -4\lambda & 1+6\lambda & -4\lambda & \lambda & \\ & & \ddots & \ddots & \ddots & \ddots & \\ & & & \lambda & -4\lambda & 1+6\lambda & -4\lambda & \lambda \\ & & & & \lambda & -4\lambda & 1+6\lambda & -4\lambda & \lambda \\ & & & & & \lambda & -4\lambda & 1+5\lambda & -2\lambda \\ & & & & & & \lambda & -2\lambda & 1+\lambda \end{bmatrix}$$

and The solution to this set of T equations in T unknowns is

$$\boldsymbol{\tau} = \Gamma^{-1} \mathbf{y}.$$

The cyclic component is defined as $c_t = y_t - \tau_t$.

Hodrick and Prescott (1997) recommend values of 100 for annual data, 1600 for quarterly data and 14400 for monthly data. The HP filter is non-causal and so is not appropriate for prediction. The gain function of the cyclic component of the HP filter with $\lambda = 1600$ is illustrated in figure 4.11. While the filter attempts to eliminate components with a frequency below ten years of quarterly data ($\frac{1}{40}$), there is some gain until about $\frac{1}{50}$ and the gain is not unity until approximately $\frac{1}{25}$.

4.12.4 Baxter-King Filter

Baxter and King (1999) consider the problem of designing a filter to be close to the ideal filter subject to using a finite number of points.¹⁸ They further argue that extracting the cyclic component requires the use of both a high-pass and a low-pass filter – the high-pass filter is to cutoff the most persistent components while the low-pass filter is used to eliminate high-frequency noise. The BK filter is defined by a triple, two-period lengths (inverse frequencies) and the number of points used to construct the filter (k), and is written as $BK_k(p, q)$ where $p < q$ are the cutoff frequencies.

Baxter and King suggest using a band-pass filter with cutoffs at $\frac{1}{32}$ and $\frac{1}{6}$ for quarterly data. The final choice for their approximate ideal filter is the number of nodes, for which they suggest 12. The

¹⁸Ideal filters, except for the trivial all-pass, require an infinite number of points to implement, and so are infeasible in applications.

number of points has two effects. First, the BK filter cannot be used in the first and last k points. Second, a higher number of nodes will produce a more accurate approximation to the ideal filter.

Implementing the BK filter is simple. Baxter and King show that the optimal weights for a low-pass filter at particular frequency f , satisfy

$$\tilde{w}_0 = 2f \quad (4.95)$$

$$\tilde{w}_i = \frac{\sin(2i\pi f)}{i\pi}, \quad i = 1, 2, \dots, k \quad (4.96)$$

$$\theta = [2k+1]^{-1} \left(1 - \sum_{i=-k}^k \tilde{w}_i \right) \quad (4.97)$$

$$w_i = \tilde{w}_i + \theta, \quad i = 0, 1, \dots, k \quad (4.98)$$

$$w_i = w_{-i}. \quad (4.99)$$

The BK filter is constructed as the difference between two low-pass filters, and so

$$\tau_t = \sum_{i=-k}^k w_i y_{t-i} \quad (4.100)$$

$$c_t = \sum_{i=-k}^k (v_i - w_i) y_{t-i} \quad (4.101)$$

where $\{w_i\}$ and $\{v_i\}$ are both weights from low-pass filters where the period used to construct $\{w_i\}$ is longer than the period used to construct $\{v_i\}$. The gain function of the $BK_{12}(6, 32)$ is illustrated in the upper right panel of figure 4.11. The approximation is reasonable, with near unit gain between $\frac{1}{32}$ and $\frac{1}{6}$ and little gain outside.

4.12.5 First Difference

Another very simple filter to separate a “trend” from a “cyclic” component is the first difference. Note that if y_t is an I(1) series, then the first difference which contains the “cyclic” component, $c_t = \frac{1}{2}\Delta y_t$, is an I(0) series and so the first difference is a causal filter. The “trend” is measured using an MA(2), $\tau_t = \frac{1}{2}(y_t + y_{t-1})$ so that $y_t = c_t + \tau_t$. The FD filter is not sharp – it allows for most frequencies to enter the cyclic component – and so is not recommended in practice.

4.12.6 Beveridge-Nelson Decomposition

The Beveridge and Nelson (1981) decomposition extends the first order difference decomposition to include any predictable component in the future trend as part of the current trend. The idea behind the BN decomposition is simple: if the predictable part of the long-run component places the long-run component above its current value, then the cyclic component should be negative. Similarly, if the predictable part of the long-run component expects that the long run component should trend lower then the cyclic component should be positive. Formally the BN decomposition if defined as

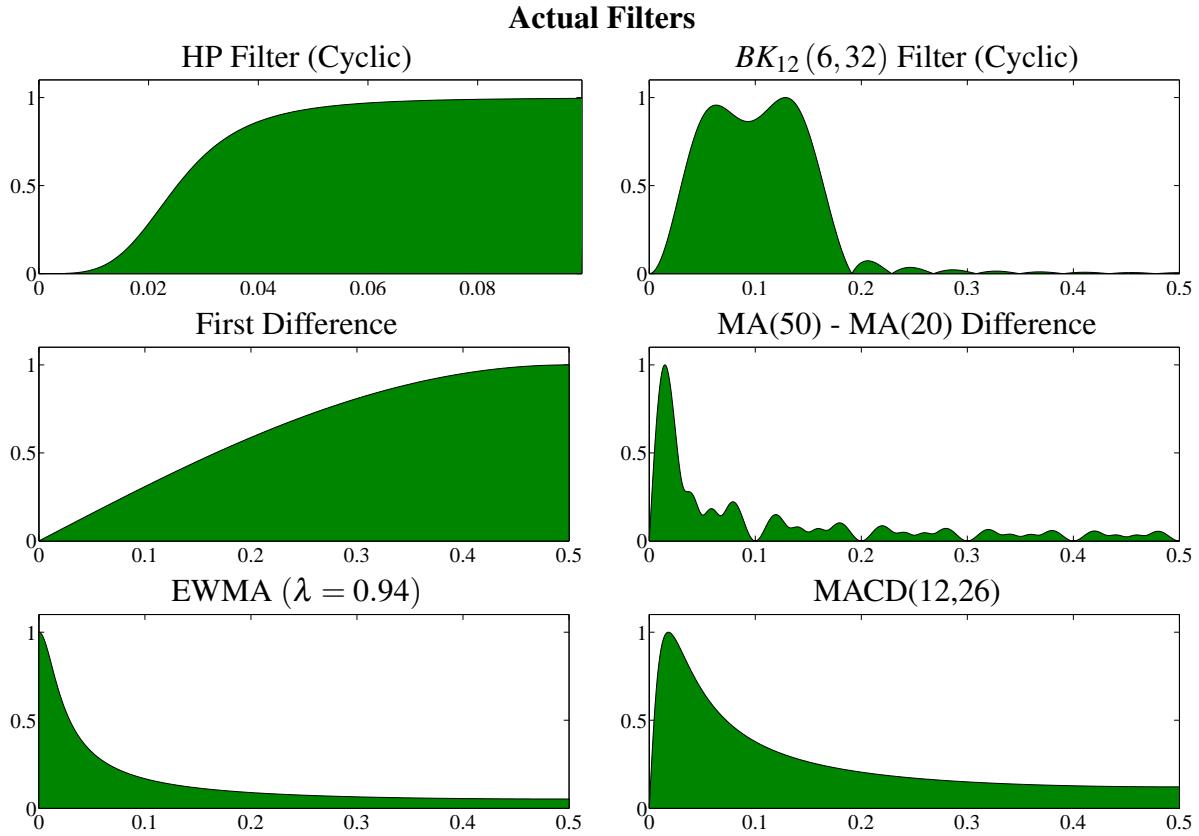


Figure 4.11: These six panels contain the standard HP filter, the $BK_{12}(6,32)$ filter, the first difference filter, an EWMA with $\lambda = .94$, a MACD(12,26) and the difference between a 20-day and a 50-day moving average. The gain functions in the right hand column have been normalized so that the maximum weight is 1. This is equivalent to scaling all of the filter weights by a constant, and so is simple a change in variance of the filter output.

$$\tau_t = \lim_{h \rightarrow \infty} \hat{y}_{t+h|t} - h\mu \quad (4.102)$$

$$c_t = y_t - \tau_t$$

where μ is the drift in the trend, if any. The trend can be equivalently expressed as the current level of y_t plus the expected increments minus the drift,

$$\tau_t = y_t + \lim_{h \rightarrow \infty} \sum_{i=1}^h E[\Delta \hat{y}_{t+i|t} - \mu] \quad (4.103)$$

where μ is the unconditional expectation of the increments to y_t , $E[\Delta \hat{y}_{t+j|t}]$. The trend component contains the persistent component and so the filter applied must be a low-pass filter, while the cyclic component is stationary and so must be the output of a high-pass filter. The gain of the filter applied when using the BN decomposition depends crucially on the forecasting model for the short-run component.

Suppose $\{y_t\}$ is an I(1) series which has both a permanent and transitive component. Since $\{y_t\}$ is I(1), Δy_t must be I(0) and so can be described by a stationary ARMA(P,Q) process. For simplicity, suppose that Δy_t follows an MA(3) so that

$$\Delta y_t = \phi_0 + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3} + \varepsilon_t$$

In this model $\mu = \phi_0$, and the h -step ahead forecast is given by

$$\begin{aligned}\hat{\Delta y}_{t+1|t} &= \mu + \theta_1 \varepsilon_t + \theta_2 \varepsilon_{t-1} + \theta_3 \varepsilon_{t-2} \\ \hat{\Delta y}_{t+2|t} &= \mu + \theta_2 \varepsilon_t + \theta_3 \varepsilon_{t-1} \\ \hat{\Delta y}_{t+3|t} &= \mu + \theta_3 \varepsilon_t \\ \hat{\Delta y}_{t+h|t} &= \mu \quad h \geq 4,\end{aligned}$$

and so

$$\tau_t = y_t + (\theta_1 + \theta_2 + \theta_3) \varepsilon_t + (\theta_2 + \theta_3) \varepsilon_{t-1} + \theta_3 \varepsilon_{t-2}$$

and

$$c_t = -(\theta_1 + \theta_2 + \theta_3) \varepsilon_t - (\theta_2 + \theta_3) \varepsilon_{t-1} - \theta_3 \varepsilon_{t-2}.$$

Alternatively, suppose that Δy_t follows an AR(1) so that

$$\Delta y_t = \phi_0 + \phi_1 \Delta y_{t-1} + \varepsilon_t.$$

This model can be equivalently defined in terms of deviations around the long-run mean, $\tilde{\Delta y}_t = \Delta y_t - \phi_0 / (1 - \phi_1)$, as

$$\begin{aligned}\Delta y_t &= \phi_0 + \phi_1 \Delta y_{t-1} + \varepsilon_t \\ \Delta y_t &= \phi_0 \frac{1 - \phi_1}{1 - \phi_1} + \phi_1 \Delta y_{t-1} + \varepsilon_t \\ \Delta y_t &= \frac{\phi_0}{1 - \phi_1} - \phi_1 \frac{\phi_0}{1 - \phi_1} + \phi_1 \Delta y_{t-1} + \varepsilon_t \\ \Delta y_t - \frac{\phi_0}{1 - \phi_1} &= \phi_1 \left(\Delta y_{t-1} - \frac{\phi_0}{1 - \phi_1} \right) + \varepsilon_t \\ \tilde{\Delta y}_t &= \phi_1 \tilde{\Delta y}_{t-1} + \varepsilon_t.\end{aligned}$$

In this transformed model, $\mu = 0$ which simplifies finding the expression for the trend. The h -step ahead forecast if $\tilde{\Delta y}_t$ is given by

$$\hat{\tilde{\Delta y}}_{t+h|t} = \phi_1^h \tilde{\Delta y}_t$$

and so

$$\begin{aligned}
\tau_t &= y_t + \lim_{h \rightarrow \infty} \sum_{i=1}^h \Delta \hat{\tilde{y}}_{t+i|t} \\
&= y_t + \lim_{h \rightarrow \infty} \sum_{i=1}^h \phi_1^i \Delta \tilde{y}_t \\
&= y_t + \lim_{h \rightarrow \infty} \Delta \tilde{y}_t \sum_{i=1}^h \phi_1^i \\
&= y_t + \lim_{h \rightarrow \infty} \Delta \tilde{y}_t \frac{\phi_1}{1 - \phi_1}
\end{aligned}$$

which follows since $\lim_{h \rightarrow \infty} \sum_{i=1}^h \phi_1^i = -1 + \lim_{h \rightarrow \infty} \sum_{i=0}^h \phi_1^i = 1/(1 - \phi_1) - 1$. The main criticism of the Beveridge-Nelson decomposition is that the trend and the cyclic component are perfectly (negatively) correlation.

4.12.7 Extracting the cyclic components from Real US GDP

To illustrate the filters, the cyclic component was extracted from log real US GDP data taken from the Federal Reserve Economics Database. Data was available from 1947 Q1 to Q2 2009. Figure 4.12 contains the cyclical component extracted using 4 methods. The top panel contains the standard HP filter with $\lambda = 1600$. The middle panel contains $BK_{12}(6, 32)$ (solid) and $BK_{12}(1, 32)$ (dashed) filters, the latter of which is a high pass-filter since the faster frequency is 1. Note that the first and last 12 points of the cyclic component are set to 0. The bottom panel contains the cyclic component extracted using a Beveridge-Nelson decomposition based on an AR(1) fit to GDP growth. For the BN decomposition, the first 2 points are zero which reflects the loss of data due to the first difference and the fitting of the AR(1) to the first difference.¹⁹

The HP filter and the $BK_{12}(1, 32)$ are remarkably similar with a correlation of over 99%. The correlation between the $BK_{12}(6, 32)$ and the HP filter was 96%, the difference being in the lack of a high-frequency component. The cyclic component from the BN decomposition has a small negative correlation with the other three filters, although choosing a different model for GDP growth would change the decomposition.

4.12.8 Markov Switching Autoregression

Markov switching autoregression, introduced into econometrics in Hamilton (1989), uses a composite model which evolves according to both an autoregression and a latent state which determines the value of the autoregressive parameters. In financial applications using low-frequency asset returns, an MSAR that allows the mean and the variance to be state-dependent has been found to outperform linear models (Perez-Quiros and Timmermann, 2000).

Definition 4.45 (Markov Switching Autoregression). A k -state Markov switching autoregression (MSAR) is a stochastic process which has dynamics that evolve through both a Markovian state

¹⁹The AR(1) was chosen from a model selection search of AR models with an order up to 8 using the SBIC.

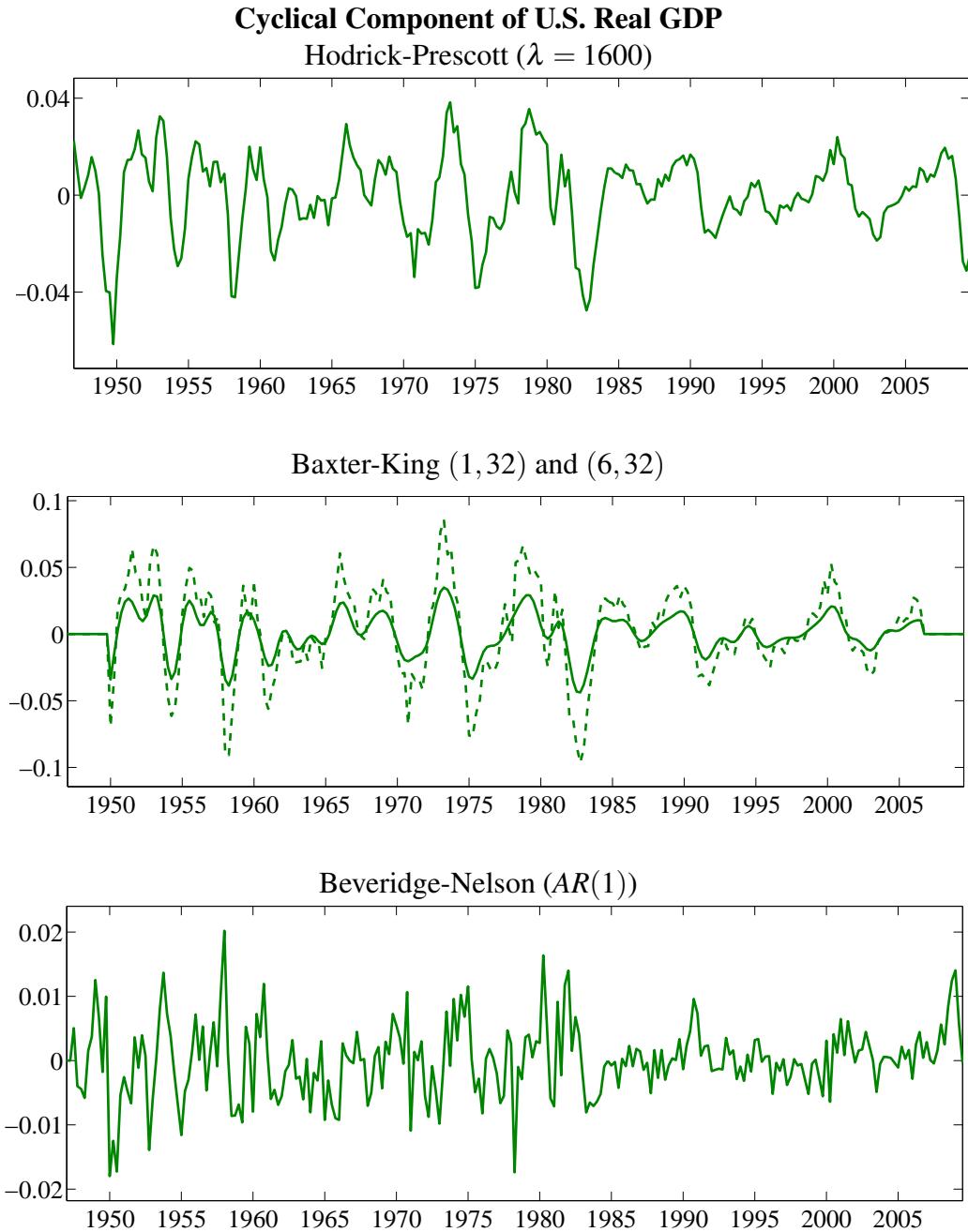


Figure 4.12: The top panel contains the filtered cyclic component from a HP filter with $\lambda = 1600$. The middle panel contains the cyclic component from $BK_{12}(6,32)$ (solid) and $BK_{12}(1,32)$ (dashed) filters. The bottom panel contains the cyclic component from a Beveridge-Nelson decomposition based on an AR(1) model for GDP growth rates.

process and an autoregression where the autoregressive parameters are state dependent. The states,

labeled $1, 2, \dots, k$, are denoted s_t and follow a k -state latent Markov Chain with transition matrix \mathbf{P} ,

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \dots & p_{1k} \\ p_{21} & p_{22} & \dots & p_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ p_{k1} & p_{k2} & \dots & p_{kk} \end{bmatrix} \quad (4.104)$$

where $p_{ij} = Pr(s_{t+1} = i | s_t = j)$. Note that the columns must sum to 1 since $\sum_{i=1}^k Pr(s_{t+1} = i | s_t = j) = 1$. Data are generated according to a P^{th} order autoregression,

$$y_t = \phi_0^{(s_t)} + \phi_1^{(s_t)} y_{t-1} + \dots + \phi_P^{(s_t)} y_{t-p} + \sigma^{(s_t)} \varepsilon_t \quad (4.105)$$

where $\phi^{(s_t)} = [\phi_0^{(s_t)} \phi_1^{(s_t)} \dots \phi_P^{(s_t)}]'$ are state-dependent autoregressive parameters, $\sigma^{(s_t)}$ is the state-dependent standard deviation and $\varepsilon_t \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$.²⁰ The unconditional state probabilities ($Pr(s_t = i)$), known as the ergodic probabilities, are denoted $\pi = [\pi_1 \pi_2 \dots \pi_k]'$ and are the solution to

$$\pi = \mathbf{P}\pi. \quad (4.106)$$

The ergodic probabilities can also be computed as the normalized eigenvector of \mathbf{P} corresponding to the only unit eigenvalue.

Rather than attempting to derive properties of an MSAR, consider a simple specification with two states, no autoregressive terms, and where only the mean of the process varies²¹

$$y_t = \begin{cases} \phi^H + \varepsilon_t \\ \phi^L + \varepsilon_t \end{cases} \quad (4.107)$$

where the two states are indexed by H (high) and L (low). The transition matrix is

$$\mathbf{P} = \begin{bmatrix} p_{HH} & p_{HL} \\ p_{LH} & p_{LL} \end{bmatrix} = \begin{bmatrix} p_{HH} & 1 - p_{LL} \\ 1 - p_{HH} & p_{LL} \end{bmatrix} \quad (4.108)$$

and the unconditional probabilities of being in the high and low state, π_H and π_L , respectively, are

$$\pi_H = \frac{1 - p_{LL}}{2 - p_{HH} - p_{LL}} \quad (4.109)$$

$$\pi_L = \frac{1 - p_{HH}}{2 - p_{HH} - p_{LL}}. \quad (4.110)$$

This simple model is useful for understanding the data generation in a Markov Switching process:

1. At $t = 0$ an initial state, s_0 , is chosen according to the ergodic (unconditional) probabilities. With probability π_H , $s_0 = H$ and with probability $\pi_L = 1 - \pi_H$, $s_0 = L$.

²⁰The assumption that $\varepsilon_t \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$ can be easily relaxed to include other i.i.d. processes for the innovations.

²¹See Hamilton (1994, chapter 22) or Krolzig (1997) for further information on implementing MSAR models.

2. The state probabilities evolve independently from the observed data according to a Markov Chain. If $s_0 = H$, $s_1 = H$ with probability p_{HH} , the probability $s_{t+1} = H$ given $s_t = H$ and $s_1 = L$ with probability $p_{LH} = 1 - p_{HH}$. If $s_0 = L$, $s_1 = H$ with probability $p_{HL} = 1 - p_{LL}$ and $s_1 = L$ with probability p_{LL} .
3. Once the state at $t = 1$ is known, the value of y_1 is chosen according to

$$y_1 = \begin{cases} \phi^H + \varepsilon_1 & \text{if } s_1 = H \\ \phi^L + \varepsilon_t & \text{if } s_1 = L \end{cases}.$$

4. Steps 2 and 3 are repeated for $t = 2, 3, \dots, T$, to produce a time series of Markov Switching data.

4.12.8.1 Markov Switching Examples

Using the 2-state Markov Switching Autoregression described above, 4 systems were simulated for 100 observations.

- Pure mixture
 - $\mu_H = 4, \mu_L = -2, V[\varepsilon_t] = 1$ in both states
 - $p_{HH} = .5 = p_{LL}$
 - $\pi_H = \pi_L = .5$
 - Remark: This is a “pure” mixture model where the probability of each state does not depend on the past. This occurs because the probability of going from high to high is the same as the probability of going from low to high, 0.5.
- Two persistent States
 - $\mu_H = 4, \mu_L = -2, V[\varepsilon_t] = 1$ in both states
 - $p_{HH} = .9 = p_{LL}$ so the average duration of each state is 10 periods.
 - $\pi_H = \pi_L = .5$
 - Remark: Unlike the first parameterization this is not a simple mixture. Conditional on the current state being H , there is a 90% chance that the next state will remain H .
- One persistent state, one transitory state
 - $\mu_H = 4, \mu_L = -2, V[\varepsilon_t] = 1$ if $s_t = H$ and $V[\varepsilon_t] = 2$ if $s_t = L$
 - $p_{HH} = .9, p_{LL} = .5$
 - $\pi_H = .83, \pi_L = .16$
 - Remark: This type of model is consistent with quarterly data on U.S. GDP where booms (H) typically last 10 quarters while recessions die quickly, typically in 2 quarters.
- Mixture with different variances

- $\mu_H = 4, \mu_L = -2, V[\varepsilon_t] = 1$ if $s_t = H$ and $V[\varepsilon_t] = 16$ if $s_t = L$
- $p_{HH} = .5 = p_{LL}$
- $\pi_H = \pi_L = .5$
- Remark: This is another “pure” mixture model but the variances differ between the states. One nice feature of mixture models (MSAR is a member of the family of mixture models) is that the unconditional distribution of the data may be non-normal even though the shocks are conditionally normally distributed.²²

Figure 4.13 contains plots of 100 data points generated from each of these processes. The first (MSAR(1)) produces a mixture with modes at -2 and 4 each with equal probability and the states (top panel, bottom right) are i.i.d.. The second process produced a similar unconditional distribution but the state evolution is very different. Each state is very persistent and, conditional on the state being high or low, it was likely to remain the same. The third process had one very persistent state and one with much less persistence. This produced a large skew in the unconditional distribution since the state where $\mu = -2$ was visited less frequently than the state with $\mu = 4$. The final process (MSAR(4)) has state dynamics similar to the first but produces a very different unconditional distribution. The difference occurs since the variance depends on the state of the Markov process.

4.12.9 Threshold Autoregression and Self-Exciting Threshold Autoregression

A second class of nonlinear models that have gained considerable traction in financial applications are Threshold Autoregressions (TAR), and in particular, the subfamily of Self-Exciting Threshold Autoregressions (SETAR).²³

Definition 4.46 (Threshold Autoregression). A threshold autoregression is a P^{th} Order autoregressive process with state-dependent parameters where the state is determined by the lagged level of an exogenous variable x_{t-k} for some $k \geq 1$.

$$y_t = \phi_0^{(s_t)} + \phi_1^{(s_t)} y_{t-1} + \dots + \phi_p^{(s_t)} y_{t-p} + \sigma^{(s_t)} \varepsilon_t \quad (4.111)$$

Let $-\infty = x_0 < x_1 < x_2 < \dots < x_N < x_{N+1} = \infty$ be a partition of x in to $N+1$ distinct bins. $s_t = j$ if $x_{t-k} \in (x_j, x_{j+1})$.

Self-exciting threshold autoregressions, introduced in Tong (1978), are similarly defined. The only change is in the definition of the threshold variable; rather than relying on an exogenous variable to determine the state, the state in SETARs is determined by lagged values of the dependent variable.

Definition 4.47 (Self Exciting Threshold Autoregression). A self exciting threshold autoregression is a P^{th} Order autoregressive process with state-dependent parameters where the state is determined by the lagged level of the dependent variable, y_{t-k} for some $k \geq 1$.

$$y_t = \phi_0^{(s_t)} + \phi_1^{(s_t)} y_{t-1} + \dots + \phi_p^{(s_t)} y_{t-p} + \sigma^{(s_t)} \varepsilon_t \quad (4.112)$$

²²Mixtures of finitely many normals, each with different means and variances, can be used approximate many non-normal distributions.

²³See Fan and Yao (2005) for a comprehensive treatment of non-linear time-series models.

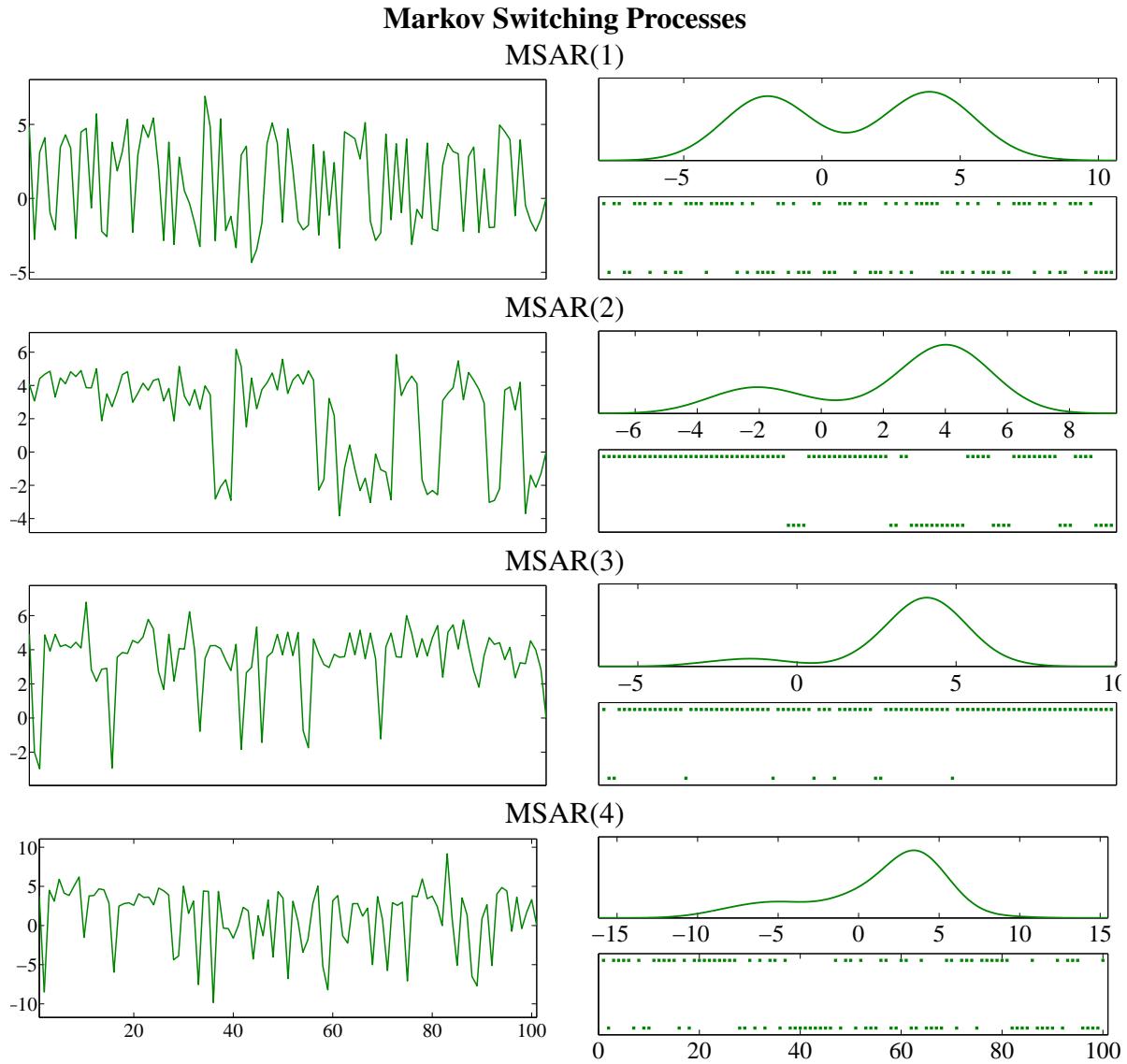


Figure 4.13: The four panels of this figure contain simulated data generated by the 4 Markov switching processes described in the text. In each panel, the large subpanel contains the generated data, the top right subpanel contains a kernel density estimate of the unconditional density and the bottom right subpanel contains the time series of the state values (high points correspond to the high state).

Let $-\infty = y_0 < y_1 < y_2 < \dots < y_N < y_{N+1} = \infty$ be a partition of y in to $N + 1$ distinct bins. $s_t = j$ is $y_{t-k} \in (y_j, y_{j+1})$.

The primary application of SETAR models in finance has been to exchange rates which often exhibit a behavior that is difficult to model with standard ARMA models: many FX rates exhibit random-walk-like behavior in a range yet remain within the band longer than would be consistent with a simple random walk. A symmetric SETAR is a parsimonious model that can describe this behavior and is parameterized

$$\begin{aligned} y_t &= y_{t-1} + \varepsilon_t \text{ if } C - \delta < y_t < C + \delta \\ y_t &= C(1 - \phi) + \phi y_{t-1} + \varepsilon_t \text{ if } y_t < C - \delta \text{ or } y_t > C + \delta \end{aligned} \quad (4.113)$$

where C is the “target” exchange rate. The first equation is a standard random walk, and when y_t is within the target band it behaves like a random walk. The second equation is only relevant when y_t is outside of its target band and ensures that y_t is mean reverting towards C as long as $|\phi| < 1$.²⁴ ϕ is usually assumed to lie between 0 and 1 which produces a smooth mean reversion back towards the band.

To illustrate the behavior of this process and the highlight the differences between it and a random walk, 200 data points were generated with different values of ϕ using standard normal innovations. The mean was set to 100 and the used $\delta = 5$, and so y_t follows a random walk when between 95 and 105. The lag value of the threshold variable (k) was set to one. Four values for ϕ were used: 0, 0.5, 0.9 and 1. The extreme cases represent a process which is immediately mean reverting ($\phi = 0$), in which case as soon as y_t leaves the target band it is immediately returned to C , and a process that is a pure random walk ($\phi = 1$) since $y_t = y_{t-1} + \varepsilon_t$ for any value of y_{t-1} . The two interior cases represent smooth reversion back to the band; when $\phi = .5$ the reversion is quick and when $\phi = .9$ the reversion is slow. When ϕ is close to 1 it is very difficult to differentiate a band SETAR from a pure random walk, which is one of the explanations for the poor performance of unit root tests where tests often fail to reject a unit root despite clear economic theory predicting that a time series should be mean reverting.

4.A Computing Autocovariance and Autocorrelations

This appendix covers the derivation of the ACF for the MA(1), MA(Q), AR(1), AR(2), AR(3) and ARMA(1,1). Throughout this appendix, $\{\varepsilon_t\}$ is assumed to be a white noise process and the processes parameters are always assumed to be consistent with covariance stationarity. All models are assumed to be mean 0, an assumption made without loss of generality since autocovariances are defined using demeaned time series,

$$\gamma_s = E[(y_t - \mu)(y_{t-s} - \mu)]$$

where $\mu = E[y_t]$. Recall that the autocorrelation is simply the s^{th} autocovariance to the variance,

$$\rho_s = \frac{\gamma_s}{\gamma_0}.$$

This appendix presents two methods for deriving the autocorrelations of ARMA processes: backward substitution and the Yule-Walker equations, a set of k equations with k unknowns where $\gamma_0, \gamma_1, \dots, \gamma_{k-1}$ are the solution.

²⁴Recall the mean of an AR(1) $y_t = \phi_0 + \phi_1 y_{t-1} + \varepsilon_t$ is $\phi_0 / (1 - \phi_1)$ where $\phi_0 = C(1 - \phi)$ and $\phi_1 = \phi$ in this SETAR.

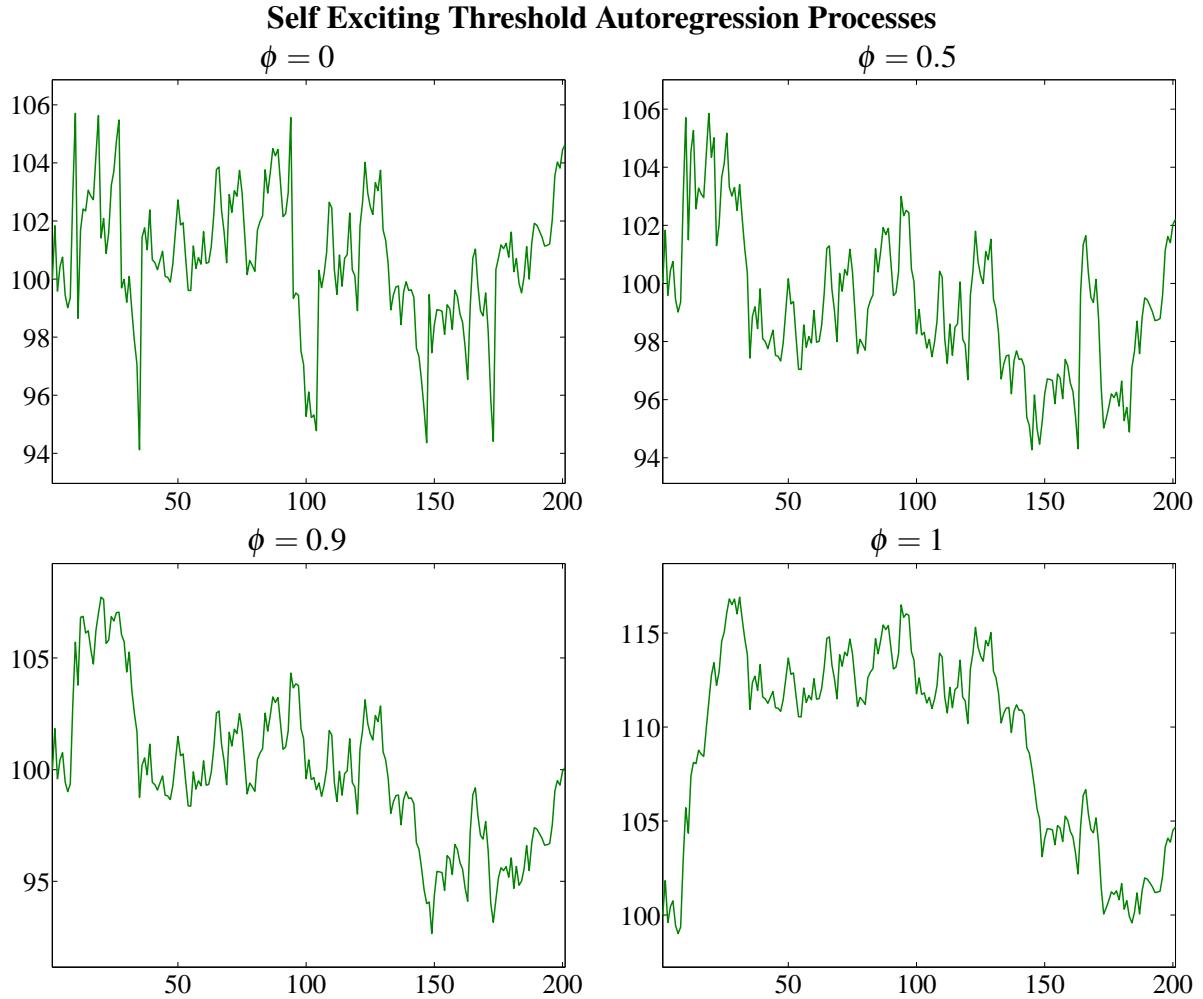


Figure 4.14: The four panels of this figure contain simulated data generated by a SETAR with different values of ϕ . When $\phi = 0$ the process is immediately returned to its unconditional mean $C = 100$. Larger values of ϕ increase the amount of time spent outside of the "target band" (95–105) and when $\phi = 1$, the process is a pure random walk.

4.A.1 Yule-Walker

The Yule-Walker equations are a linear system of $\max(P, Q) + 1$ equations (in an ARMA(P,Q)) where the solution to the system are the long-run variance and the first $k - 1$ autocovariances. The Yule-Walker equations are formed by equating the definition of an autocovariance with an expansion produced by substituting for the contemporaneous value of y_t . For example, suppose y_t follows an AR(2) process,

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$$

The variance must satisfy

$$\text{E}[y_t y_t] = \text{E}[y_t (\phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t)] \quad (4.114)$$

$$\begin{aligned} E[y_t^2] &= E[\phi_1 y_t y_{t-1} + \phi_2 y_t y_{t-2} + y_t \varepsilon_t] \\ V[y_t] &= \phi_1 E[y_t y_{t-1}] + \phi_2 E[y_t y_{t-2}] + E[y_t \varepsilon_t]. \end{aligned}$$

In the final equation above, terms of the form $E[y_t y_{t-s}]$ are replaced by their population values, γ_s and $E[y_t \varepsilon_t]$ is replaced with its population value, σ^2 .

$$V[y_t y_t] = \phi_1 E[y_t y_{t-1}] + \phi_2 E[y_t y_{t-2}] + E[y_t \varepsilon_t] \quad (4.115)$$

becomes

$$\gamma_0 = \phi_1 \gamma_1 + \phi_2 \gamma_2 + \sigma^2 \quad (4.116)$$

and so the long run variance is a function of the first two autocovariances, the model parameters, and the innovation variance. This can be repeated for the first autocovariance,

$$E[y_t y_{t-1}] = \phi_1 E[y_{t-1} y_{t-1}] + \phi_2 E[y_{t-1} y_{t-2}] + E[y_{t-1} \varepsilon_t]$$

becomes

$$\gamma_1 = \phi_1 \gamma_0 + \phi_2 \gamma_1, \quad (4.117)$$

and for the second autocovariance,

$$E[y_t y_{t-2}] = \phi_1 E[y_{t-2} y_{t-1}] + \phi_2 E[y_{t-2} y_{t-2}] + E[y_{t-2} \varepsilon_t] \text{ becomes}$$

becomes

$$\gamma_2 = \phi_1 \gamma_1 + \phi_2 \gamma_0. \quad (4.118)$$

Together eqs. (4.116), (4.117) and (4.118) form a system of three equations with three unknowns. The Yule-Walker method relies heavily on the covariance stationarity and so $E[y_t y_{t-j}] = E[y_{t-h} y_{t-h-j}]$ for any h . This property of covariance stationary processes was repeatedly used in forming the producing the Yule-Walker equations since $E[y_t y_t] = E[y_{t-1} y_{t-1}] = E[y_{t-2} y_{t-2}] = \gamma_0$ and $E[y_t y_{t-1}] = E[y_{t-1} y_{t-2}] = \gamma_1$.

The Yule-Walker method will be demonstrated for a number of models, starting from a simple MA(1) and working up to an ARMA(1,1).

4.A.2 MA(1)

The MA(1) is the simplest model to work with.

$$y_t = \theta_1 \varepsilon_{t-1} + \varepsilon_t$$

The Yule-Walker equation are

$$\begin{aligned} E[y_t y_t] &= E[\theta_1 \varepsilon_{t-1} y_t] + E[\varepsilon_t y_t] \\ E[y_t y_{t-1}] &= E[\theta_1 \varepsilon_{t-1} y_{t-1}] + E[\varepsilon_t y_{t-1}] \\ E[y_t y_{t-2}] &= E[\theta_1 \varepsilon_{t-1} y_{t-2}] + E[\varepsilon_t y_{t-2}] \end{aligned} \quad (4.119)$$

$$\begin{aligned} \gamma_0 &= \theta_1^2 \sigma^2 + \sigma^2 \\ \gamma_1 &= \theta_1 \sigma^2 \\ \gamma_2 &= 0 \end{aligned} \quad (4.120)$$

Additionally, both γ_s and ρ_s , $s \geq 2$ are 0 by the white noise property of the residuals, and so the autocorrelations are

$$\begin{aligned} \rho_1 &= \frac{\theta_1 \sigma^2}{\theta_1^2 \sigma^2 + \sigma^2} \\ &= \frac{\theta_1}{1 + \theta_1^2}, \\ \rho_2 &= 0. \end{aligned}$$

4.A.2.1 MA(Q)

The Yule-Walker equations can be constructed and solved for any MA(Q), and the structure of the autocovariance is simple to detect by constructing a subset of the full system.

$$\begin{aligned} E[y_t y_t] &= E[\theta_1 \varepsilon_{t-1} y_t] + E[\theta_2 \varepsilon_{t-2} y_t] + E[\theta_3 \varepsilon_{t-3} y_t] + \dots + E[\theta_Q \varepsilon_{t-Q} y_t] \\ \gamma_0 &= \theta_1^2 \sigma^2 + \theta_2^2 \sigma^2 + \theta_3^2 \sigma^2 + \dots + \theta_Q^2 \sigma^2 + \sigma^2 \\ &= \sigma^2(1 + \theta_1^2 + \theta_2^2 + \theta_3^2 + \dots + \theta_Q^2) \end{aligned} \quad (4.121)$$

$$\begin{aligned} E[y_t y_{t-1}] &= E[\theta_1 \varepsilon_{t-1} y_{t-1}] + E[\theta_2 \varepsilon_{t-2} y_{t-1}] + E[\theta_3 \varepsilon_{t-3} y_{t-1}] + \dots + E[\theta_Q \varepsilon_{t-Q} y_{t-1}] \\ \gamma_1 &= \theta_1 \sigma^2 + \theta_1 \theta_2 \sigma^2 + \theta_2 \theta_3 \sigma^2 + \dots + \theta_{Q-1} \theta_Q \sigma^2 \\ &= \sigma^2(\theta_1 + \theta_1 \theta_2 + \theta_2 \theta_3 + \dots + \theta_{Q-1} \theta_Q) \end{aligned} \quad (4.122)$$

$$\begin{aligned} E[y_t y_{t-2}] &= E[\theta_1 \varepsilon_{t-1} y_{t-2}] + E[\theta_2 \varepsilon_{t-2} y_{t-2}] + E[\theta_3 \varepsilon_{t-3} y_{t-2}] + \dots + E[\theta_Q \varepsilon_{t-Q} y_{t-2}] \\ \gamma_2 &= \theta_2 \sigma^2 + \theta_1 \theta_3 \sigma^2 + \theta_2 \theta_4 \sigma^2 + \dots + \theta_{Q-2} \theta_Q \sigma^2 \\ &= \sigma^2(\theta_2 + \theta_1 \theta_3 + \theta_2 \theta_4 + \dots + \theta_{Q-2} \theta_Q) \end{aligned} \quad (4.123)$$

The pattern that emerges shows,

$$\gamma_s = \theta_s \sigma^2 + \sum_{i=1}^{Q-s} \sigma^2 \theta_i \theta_{i+s} = \sigma^2 (\theta_s + \sum_{i=1}^{Q-s} \theta_i \theta_{i+s}).$$

and so, γ_s is a sum of $Q - s + 1$ terms. The autocorrelations are

$$\begin{aligned}\rho_1 &= \frac{\theta_1 + \sum_{i=1}^{Q-1} \theta_i \theta_{i+1}}{1 + \theta_s + \sum_{i=1}^Q \theta_i^2} \\ \rho_2 &= \frac{\theta_2 + \sum_{i=1}^{Q-2} \theta_i \theta_{i+2}}{1 + \theta_s + \sum_{i=1}^Q \theta_i^2} \\ &\vdots = \vdots \\ \rho_Q &= \frac{\theta_Q}{1 + \theta_s + \sum_{i=1}^Q \theta_i^2} \\ \rho_{Q+s} &= 0, \quad s \geq 0\end{aligned}\tag{4.124}$$

4.A.2.2 AR(1)

The Yule-Walker method requires $\max(P, Q) + 1$ equations to compute the autocovariance for an ARMA(P, Q) process and in an AR(1), two are required (the third is included to establish this point).

$$y_t = \phi_1 y_{t-1} + \varepsilon_t$$

$$\begin{aligned}E[y_t y_t] &= E[\phi_1 y_{t-1} y_t] + E[\varepsilon_t y_t] \\ E[y_t y_{t-1}] &= E[\phi_1 y_{t-1} y_{t-1}] + E[\varepsilon_t y_{t-1}] \\ E[y_t y_{t-2}] &= E[\phi_1 y_{t-1} y_{t-2}] + E[\varepsilon_t y_{t-2}]\end{aligned}\tag{4.125}$$

These equations can be rewritten in terms of the autocovariances, model parameters and σ^2 by taking expectation and noting that $E[\varepsilon_t y_t] = \sigma^2$ since $y_t = \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_1^2 \varepsilon_{t-2} + \dots$ and $E[\varepsilon_t y_{t-j}] = 0$, $j > 0$ since $\{\varepsilon_t\}$ is a white noise process.

$$\begin{aligned}\gamma_0 &= \phi_1 \gamma_1 + \sigma^2 \\ \gamma_1 &= \phi_1 \gamma_0 \\ \gamma_2 &= \phi_1 \gamma_1\end{aligned}\tag{4.126}$$

The third is redundant since γ_2 is fully determined by γ_1 and ϕ_1 , and higher autocovariances are similarly redundant since $\gamma_s = \phi_1 \gamma_{s-1}$ for any s . The first two equations can be solved for γ_0 and γ_1 ,

$$\begin{aligned}\gamma_0 &= \phi_1 \gamma_1 + \sigma^2 \\ \gamma_1 &= \phi_1 \gamma_0 \\ \Rightarrow \gamma_0 &= \phi_1^2 \gamma_0 + \sigma^2 \\ \Rightarrow \gamma_0 - \phi_1^2 \gamma_0 &= \sigma^2 \\ \Rightarrow \gamma_0(1 - \phi_1^2) &= \sigma^2 \\ \Rightarrow \gamma_0 &= \frac{\sigma^2}{1 - \phi_1^2}\end{aligned}$$

and

$$\begin{aligned}\gamma_1 &= \phi_1 \gamma_0 \\ \gamma_0 &= \frac{\sigma^2}{1 - \phi_1^2} \\ \Rightarrow \gamma_1 &= \phi_1 \frac{\sigma^2}{1 - \phi_1^2}.\end{aligned}$$

The remaining autocovariances can be computed using the recursion $\gamma_s = \phi_1 \gamma_{s-1}$, and so

$$\gamma_s = \phi_1^2 \frac{\sigma^2}{1 - \phi_1^2}.$$

Finally, the autocorrelations can be computed as ratios of autocovariances,

$$\begin{aligned}\rho_1 &= \frac{\gamma_1}{\gamma_0} = \phi_1 \frac{\sigma^2}{1 - \phi_1^2} / \frac{\sigma^2}{1 - \phi_1^2} \\ \rho_1 &= \phi_1 \\ \rho_s &= \frac{\gamma_s}{\gamma_0} = \phi_1^s \frac{\sigma^2}{1 - \phi_1^2} / \frac{\sigma^2}{1 - \phi_1^2} \\ \rho_s &= \phi_1^s.\end{aligned}$$

4.A.2.3 AR(2)

The autocorrelations in an AR(2)

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$$

can be similarly computed using the $\max(P, Q) + 1$ equation Yule-Walker system,

$$\begin{aligned}E[y_t y_t] &= \phi_1 E[y_{t-1} y_t] + \phi_2 E[y_{t-2} y_t] + E[\varepsilon_t y_t] \\ E[y_t y_{t-1}] &= \phi_1 E[y_{t-1} y_{t-1}] + \phi_2 E[y_{t-2} y_{t-1}] + E[\varepsilon_t y_{t-1}] \\ E[y_t y_{t-2}] &= \phi_1 E[y_{t-1} y_{t-2}] + \phi_2 E[y_{t-2} y_{t-2}] + E[\varepsilon_t y_{t-2}]\end{aligned}\tag{4.127}$$

and then replacing expectations with their population counterparts, $\gamma_0, \gamma_1, \gamma_2$ and σ^2 .

$$\begin{aligned}\gamma_0 &= \phi_1 \gamma_1 + \phi_2 \gamma_2 + \sigma^2 \\ \gamma_1 &= \phi_1 \gamma_0 + \phi_2 \gamma_1 \\ \gamma_2 &= \phi_1 \gamma_1 + \phi_2 \gamma_0\end{aligned}\tag{4.128}$$

Further, it must be the case that $\gamma_s = \phi_1\gamma_{s-1} + \phi_2\gamma_{s-2}$ for $s \geq 2$. To solve this system of equations, divide the autocovariance equations by γ_0 , the long run variance. Omitting the first equation, the system reduces to two equations in two unknowns,

$$\begin{aligned}\rho_1 &= \phi_1\rho_0 + \phi_2\rho_1 \\ \rho_2 &= \phi_1\rho_1 + \phi_2\rho_0\end{aligned}$$

since $\rho_0 = \gamma_0/\gamma_0 = 1$.

$$\begin{aligned}\rho_1 &= \phi_1 + \phi_2\rho_1 \\ \rho_2 &= \phi_1\rho_1 + \phi_2\end{aligned}$$

Solving this system,

$$\begin{aligned}\rho_1 &= \phi_1 + \phi_2\rho_1 \\ \rho_1 - \phi_2\rho_1 &= \phi_1 \\ \rho_1(1 - \phi_2) &= \phi_1 \\ \rho_1 &= \frac{\phi_1}{1 - \phi_2}\end{aligned}$$

and

$$\begin{aligned}\rho_2 &= \phi_1\rho_1 + \phi_2 \\ &= \phi_1 \frac{\phi_1}{1 - \phi_2} + \phi_2 \\ &= \frac{\phi_1\phi_1 + (1 - \phi_2)\phi_2}{1 - \phi_2} \\ &= \frac{\phi_1^2 + \phi_2 - \phi_2^2}{1 - \phi_2}\end{aligned}$$

Since $\rho_s = \phi_1\rho_{s-1} + \phi_2\rho_{s-2}$, these first two autocorrelations are sufficient to compute the other autocorrelations,

$$\begin{aligned}\rho_3 &= \phi_1\rho_2 + \phi_2\rho_1 \\ &= \phi_1 \frac{\phi_1^2 + \phi_2 - \phi_2^2}{1 - \phi_2} + \phi_2 \frac{\phi_1}{1 - \phi_2}\end{aligned}$$

and the long run variance of y_t ,

$$\begin{aligned}\gamma_0 &= \phi_1\gamma_1 + \phi_2\gamma_2 + \sigma^2 \\ \gamma_0 - \phi_1\gamma_1 - \phi_2\gamma_2 &= \sigma^2 \\ \gamma_0(1 - \phi_1\rho_1 - \phi_2\rho_2) &= \sigma^2 \\ \gamma_0 &= \frac{\sigma^2}{1 - \phi_1\rho_1 - \phi_2\rho_2}\end{aligned}$$

The final solution is computed by substituting for ρ_1 and ρ_2 ,

$$\begin{aligned}\gamma_0 &= \frac{\sigma^2}{1 - \phi_1 \frac{\phi_1}{1-\phi_2} - \phi_2 \frac{\phi_1^2 + \phi_2 - \phi_2^2}{1-\phi_2}} \\ &= \frac{1 - \phi_2}{1 + \phi_2} \left(\frac{\sigma^2}{(\phi_1 + \phi_2 - 1)(\phi_2 - \phi_1 - 1)} \right)\end{aligned}$$

4.A.2.4 AR(3)

Begin by constructing the Yule-Walker equations,

$$\begin{aligned}E[y_t y_t] &= \phi_1 E[y_{t-1} y_t] + \phi_2 E[y_{t-2} y_t] + \phi_3 E[y_{t-3} y_t] + E[\varepsilon_t y_t] \\ E[y_t y_{t-1}] &= \phi_1 E[y_{t-1} y_{t-1}] + \phi_2 E[y_{t-2} y_{t-1}] + \phi_3 E[y_{t-3} y_{t-1}] + E[\varepsilon_t y_{t-1}] \\ E[y_t y_{t-2}] &= \phi_1 E[y_{t-1} y_{t-2}] + \phi_2 E[y_{t-2} y_{t-2}] + \phi_3 E[y_{t-3} y_{t-2}] + E[\varepsilon_t y_{t-2}] \\ E[y_t y_{t-3}] &= \phi_1 E[y_{t-1} y_{t-3}] + \phi_2 E[y_{t-2} y_{t-3}] + \phi_3 E[y_{t-3} y_{t-3}] + E[\varepsilon_t y_{t-4}].\end{aligned}$$

Replacing the expectations with their population values, $\gamma_0, \gamma_1, \dots$ and σ^2 , the Yule-Walker equations can be rewritten

$$\begin{aligned}\gamma_0 &= \phi_1\gamma_1 + \phi_2\gamma_2 + \phi_3\gamma_3 + \sigma^2 \\ \gamma_1 &= \phi_1\gamma_0 + \phi_2\gamma_1 + \phi_3\gamma_2 \\ \gamma_2 &= \phi_1\gamma_1 + \phi_2\gamma_0 + \phi_3\gamma_1 \\ \gamma_3 &= \phi_1\gamma_2 + \phi_2\gamma_1 + \phi_3\gamma_0\end{aligned}\tag{4.129}$$

and the recursive relationship $\gamma_s = \phi_1\gamma_{s-1} + \phi_2\gamma_{s-2} + \phi_3\gamma_{s-3}$ can be observed for $s \geq 3$.

Omitting the first condition and dividing by γ_0 ,

$$\begin{aligned}\rho_1 &= \phi_1\rho_0 + \phi_2\rho_1 + \phi_3\rho_2 \\ \rho_2 &= \phi_1\rho_1 + \phi_2\rho_0 + \phi_3\rho_1 \\ \rho_3 &= \phi_1\rho_2 + \phi_2\rho_1 + \phi_3\rho_0.\end{aligned}$$

leaving three equations in three unknowns since $\rho_0 = \gamma_0/\gamma_0 = 1$.

$$\begin{aligned}\rho_1 &= \phi_1 + \phi_2\rho_1 + \phi_3\rho_2 \\ \rho_2 &= \phi_1\rho_1 + \phi_2 + \phi_3\rho_1 \\ \rho_3 &= \phi_1\rho_2 + \phi_2\rho_1 + \phi_3\end{aligned}$$

Following some tedious algebra, the solution to this system is

$$\begin{aligned}\rho_1 &= \frac{\phi_1 + \phi_2\phi_3}{1 - \phi_2 - \phi_1\phi_3 - \phi_3^2} \\ \rho_2 &= \frac{\phi_2 + \phi_1^2 + \phi_3\phi_1 - \phi_2^2}{1 - \phi_2 - \phi_1\phi_3 - \phi_3^2} \\ \rho_3 &= \frac{\phi_3 + \phi_1^3 + \phi_1^2\phi_3 + \phi_1\phi_2^2 + 2\phi_1\phi_2 + \phi_2^2\phi_3 - \phi_2\phi_3 - \phi_1\phi_3^2 - \phi_3^3}{1 - \phi_2 - \phi_1\phi_3 - \phi_3^2}.\end{aligned}$$

Finally, the unconditional variance can be computed using the first three autocorrelations,

$$\begin{aligned}\gamma_0 &= \phi_1\gamma_1 + \phi_2\gamma_2 + \phi_3\gamma_3\sigma^2 \\ \gamma_0 - \phi_1\gamma_1 - \phi_2\gamma_2 - \phi_3\gamma_3 &= \sigma^2 \\ \gamma_0(1 - \phi_1\rho_1 + \phi_2\rho_2 + \phi_3\rho_3) &= \sigma^2 \\ \gamma_0 &= \frac{\sigma^2}{1 - \phi_1\rho_1 - \phi_2\rho_2 - \phi_3\rho_3} \\ \gamma_0 &= \frac{\sigma^2(1 - \phi_2 - \phi_1\phi_3 - \phi_3^2)}{(1 - \phi_2 - \phi_3 - \phi_1)(1 + \phi_2 + \phi_3\phi_1 - \phi_3^2)(1 + \phi_3 + \phi_1 - \phi_2)}\end{aligned}$$

4.A.2.5 ARMA(1,1)

Deriving the autocovariances and autocorrelations of an ARMA process is slightly more difficult than for a pure AR or MA process. An ARMA(1,1) is specified as

$$y_t = \phi_1 y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t$$

and since $P = Q = 1$, the Yule-Walker system requires two equations, noting that the third or higher autocovariance is a trivial function of the first two autocovariances.

$$\begin{aligned}E[y_t y_t] &= E[\phi_1 y_{t-1} y_t] + E[\theta_1 \varepsilon_{t-1} y_t] + E[\varepsilon_t y_t] \\ E[y_t y_{t-1}] &= E[\phi_1 y_{t-1} y_{t-1}] + E[\theta_1 \varepsilon_{t-1} y_{t-1}] + E[\varepsilon_t y_{t-1}]\end{aligned}\tag{4.130}$$

The presence of the $E[\theta_1 \varepsilon_{t-1} y_t]$ term in the first equation complicates solving this system since ε_{t-1} appears in y_t directly through $\theta_1 \varepsilon_{t-1}$ and indirectly through $\phi_1 y_{t-1}$. The non-zero relationships can be determined by recursively substituting y_t until it consists of only ε_t , ε_{t-1} and y_{t-2} (since y_{t-2} is uncorrelated with ε_{t-1} by the WN assumption).

$$\begin{aligned}
y_t &= \phi_1 y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t \\
&= \phi_1(\phi_1 y_{t-2} + \theta_1 \varepsilon_{t-2} + \varepsilon_{t-1}) + \theta_1 \varepsilon_{t-1} + \varepsilon_t \\
&= \phi_1^2 y_{t-2} + \phi_1 \theta_1 \varepsilon_{t-2} + \phi_1 \varepsilon_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t \\
&= \phi_1^2 y_{t-2} + \phi_1 \theta_1 \varepsilon_{t-2} + (\phi_1 + \theta_1) \varepsilon_{t-1} + \varepsilon_t
\end{aligned} \tag{4.131}$$

and so $E[\theta_1 \varepsilon_{t-1} y_t] = \theta_1(\phi_1 + \theta_1)\sigma^2$ and the Yule-Walker equations can be expressed using the population moments and model parameters.

$$\begin{aligned}
\gamma_0 &= \phi_1 \gamma_1 + \theta_1(\phi_1 + \theta_1)\sigma^2 + \sigma^2 \\
\gamma_1 &= \phi_1 \gamma_0 + \theta_1 \sigma^2
\end{aligned}$$

These two equations in two unknowns which can be solved,

$$\begin{aligned}
\gamma_0 &= \phi_1 \gamma_1 + \theta_1(\phi_1 + \theta_1)\sigma^2 + \sigma^2 \\
&= \phi_1(\phi_1 \gamma_0 + \theta_1 \sigma^2) + \theta_1(\phi_1 + \theta_1)\sigma^2 + \sigma^2 \\
&= \phi_1^2 \gamma_0 + \phi_1 \theta_1 \sigma^2 + \theta_1(\phi_1 + \theta_1)\sigma^2 + \sigma^2 \\
\gamma_0 - \phi_1^2 \gamma_0 &= \sigma^2(\phi_1 \theta_1 + \phi_1 \theta_1 + \theta_1^2 + 1) \\
\gamma_0 &= \frac{\sigma^2(1 + \theta_1^2 + 2\phi_1 \theta_1)}{1 - \phi_1^2}
\end{aligned}$$

$$\begin{aligned}
\gamma_1 &= \phi_1 \gamma_0 + \theta_1 \sigma^2 \\
&= \phi_1 \left(\frac{\sigma^2(1 + \theta_1^2 + 2\phi_1 \theta_1)}{1 - \phi_1^2} \right) + \theta_1 \sigma^2 \\
&= \phi_1 \left(\frac{\sigma^2(1 + \theta_1^2 + 2\phi_1 \theta_1)}{1 - \phi_1^2} \right) + \frac{(1 - \phi_1^2)\theta_1 \sigma^2}{1 - \phi_1^2} \\
&= \frac{\sigma^2(\phi_1 + \phi_1 \theta_1^2 + 2\phi_1^2 \theta_1)}{1 - \phi_1^2} + \frac{(\theta_1 - \theta_1 \phi_1^2)\sigma^2}{1 - \phi_1^2} \\
&= \frac{\sigma^2(\phi_1 + \phi_1 \theta_1^2 + 2\phi_1^2 \theta_1 + \theta_1 - \phi_1^2 \theta_1)}{1 - \phi_1^2} \\
&= \frac{\sigma^2(\phi_1^2 \theta_1 + \phi_1 \theta_1^2 + \phi_1 + \theta_1)}{1 - \phi_1^2} \\
&= \frac{\sigma^2(\phi_1 + \theta_1)(\phi_1 \theta_1 + 1)}{1 - \phi_1^2}
\end{aligned}$$

and so the 1st autocorrelation is

$$\rho_1 = \frac{\frac{\sigma^2(\phi_1 + \theta_1)(\phi_1 \theta_1 + 1)}{1 - \phi_1^2}}{\frac{\sigma^2(1 + \theta_1^2 + 2\phi_1 \theta_1)}{1 - \phi_1^2}} = \frac{(\phi_1 + \theta_1)(\phi_1 \theta_1 + 1)}{(1 + \theta_1^2 + 2\phi_1 \theta_1)}.$$

Returning to the next Yule-Walker equation,

$$E[y_t y_{t-2}] = E[\phi_1 y_{t-1} y_{t-2}] + E[\theta_1 \varepsilon_{t-1} y_{t-2}] + E[\varepsilon_t y_{t-2}]$$

and so $\gamma_2 = \phi_1 \gamma_1$, and, dividing both sides by γ_0 , $\rho_2 = \phi_1 \rho_1$. Higher order autocovariances and autocorrelation follow $\gamma_s = \phi_1 \gamma_{s-1}$ and $\rho_s = \phi_1 \rho_{s-1}$ respectively, and so $\rho_s = \phi_1^{s-1} \rho_1$, $s \geq 2$.

4.A.3 Backward Substitution

Backward substitution is a direct but tedious method to derive the ACF and long run variance.

4.A.3.1 AR(1)

The AR(1) process,

$$y_t = \phi_1 y_{t-1} + \varepsilon_t$$

is stationary if $|\phi_1| < 1$ and $\{\varepsilon_t\}$ is white noise. To compute the autocovariances and autocorrelations using backward substitution, $y_t = \phi_1 y_{t-1} + \varepsilon_t$ must be transformed into a pure MA process by recursive substitution,

$$\begin{aligned} y_t &= \phi_1 y_{t-1} + \varepsilon_t & (4.132) \\ &= \phi_1(\phi_1 y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\ &= \phi_1^2 y_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t \\ &= \phi_1^2(\phi_1 y_{t-3} + \varepsilon_{t-2}) + \phi_1 \varepsilon_{t-1} + \varepsilon_t \\ &= \phi_1^3 y_{t-3} + \phi_1^2 \varepsilon_{t-2} + \phi_1 \varepsilon_{t-1} + \varepsilon_t \\ &= \varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_1^2 \varepsilon_{t-2} + \phi_1^3 \varepsilon_{t-3} + \dots \\ y_t &= \sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-i}. \end{aligned}$$

The variance is the expectation of the square,

$$\begin{aligned} \gamma_0 &= V[y_t] = E[y_t^2] & (4.133) \\ &= E[(\sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-i})^2] \\ &= E[(\varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_1^2 \varepsilon_{t-2} + \phi_1^3 \varepsilon_{t-3} + \dots)^2] \\ &= E[\sum_{i=0}^{\infty} \phi_1^{2i} \varepsilon_{t-i}^2 + \sum_{i=0}^{\infty} \sum_{j=0, i \neq j}^{\infty} \phi_1^i \phi_1^j \varepsilon_{t-i} \varepsilon_{t-j}] \\ &= E[\sum_{i=0}^{\infty} \phi_1^{2i} \varepsilon_{t-i}^2] + E[\sum_{i=0}^{\infty} \sum_{j=0, i \neq j}^{\infty} \phi_1^i \phi_1^j \varepsilon_{t-i} \varepsilon_{t-j}] \end{aligned}$$

$$\begin{aligned}
&= \sum_{i=0}^{\infty} \phi_1^{2i} E[\varepsilon_{t-i}^2] + \sum_{i=0}^{\infty} \sum_{j=0, i \neq j}^{\infty} \phi_1^i \phi_1^j E[\varepsilon_{t-i} \varepsilon_{t-j}] \\
&= \sum_{i=0}^{\infty} \phi_1^{2i} \sigma^2 + \sum_{i=0}^{\infty} \sum_{j=0, i \neq j}^{\infty} \phi_1^i \phi_1^j 0 \\
&= \sum_{i=0}^{\infty} \phi_1^{2i} \sigma^2 \\
&= \frac{\sigma^2}{1 - \phi_1^{2i}}
\end{aligned}$$

The difficult step in the derivation is splitting up the ε_{t-i} into those that are matched to their own lag (ε_{t-i}^2) to those which are not ($\varepsilon_{t-i} \varepsilon_{t-j}$, $i \neq j$). The remainder of the derivation follows from the assumption that $\{\varepsilon_t\}$ is a white noise process, and so $E[\varepsilon_{t-i}^2] = \sigma^2$ and $E[\varepsilon_{t-i} \varepsilon_{t-j}] = 0$, $i \neq j$. Finally, the identity that $\lim_{n \rightarrow \infty} \sum_{i=0}^n \phi_1^{2i} = \lim_{n \rightarrow \infty} \sum_{i=0}^n (\phi_1^2)^i = \frac{1}{1 - \phi_1^2}$ for $|\phi_1| < 1$ was used to simplify the expression.

The 1st autocovariance can be computed using the same steps on the MA(∞) representation,

$$\begin{aligned}
\gamma_1 &= E[y_t y_{t-1}] \tag{4.134} \\
&= E\left[\sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-i} \sum_{i=1}^{\infty} \phi_1^{i-1} \varepsilon_{t-i}\right] \\
&= E[(\varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_1^2 \varepsilon_{t-2} + \phi_1^3 \varepsilon_{t-3} + \dots)(\varepsilon_{t-1} + \phi_1 \varepsilon_{t-2} + \phi_1^2 \varepsilon_{t-3} + \phi_1^3 \varepsilon_{t-4} + \dots)] \\
&= E\left[\sum_{i=0}^{\infty} \phi_1^{2i+1} \varepsilon_{t-1-i}^2 + \sum_{i=0}^{\infty} \sum_{j=1, i \neq j}^{\infty} \phi_1^i \phi_1^{j-1} \varepsilon_{t-i} \varepsilon_{t-j}\right] \\
&= E\left[\phi_1 \sum_{i=0}^{\infty} \phi_1^{2i} \varepsilon_{t-1-i}^2\right] + E\left[\sum_{i=0}^{\infty} \sum_{j=1, i \neq j}^{\infty} \phi_1^i \phi_1^{j-1} \varepsilon_{t-i} \varepsilon_{t-j}\right] \\
&= \phi_1 \sum_{i=0}^{\infty} \phi_1^{2i} E[\varepsilon_{t-1-i}^2] + \sum_{i=0}^{\infty} \sum_{j=1, i \neq j}^{\infty} \phi_1^i \phi_1^{j-1} E[\varepsilon_{t-i} \varepsilon_{t-j}] \\
&= \phi_1 \sum_{i=0}^{\infty} \phi_1^{2i} \sigma^2 + \sum_{i=0}^{\infty} \sum_{j=1, i \neq j}^{\infty} \phi_1^i \phi_1^{j-1} 0 \\
&= \phi_1 \left(\sum_{i=0}^{\infty} \phi_1^{2i} \sigma^2 \right) \\
&= \phi_1 \frac{\sigma^2}{1 - \phi_1^2} \\
&= \phi_1 \gamma_0
\end{aligned}$$

and the s^{th} autocovariance can be similarly determined.

$$\begin{aligned}
\gamma_s &= E[y_t y_{t-s}] \tag{4.135} \\
&= E\left[\sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-i} \sum_{i=s}^{\infty} \phi_1^{i-s} \varepsilon_{t-i}\right] \\
&= E\left[\sum_{i=0}^{\infty} \phi_1^{2i+s} \varepsilon_{t-s-i}^2 + \sum_{i=0}^{\infty} \sum_{j=s, i \neq j}^{\infty} \phi_1^i \phi_1^{j-s} \varepsilon_{t-i} \varepsilon_{t-j}\right] \\
&= E\left[\phi_1^s \sum_{i=0}^{\infty} \phi_1^{2i} \varepsilon_{t-s-i}^2\right] + E\left[\sum_{i=0}^{\infty} \sum_{j=s, i \neq j}^{\infty} \phi_1^i \phi_1^{j-s} \varepsilon_{t-i} \varepsilon_{t-j}\right] \\
&= \phi_1^s \sum_{i=0}^{\infty} \phi_1^{2i} \sigma^2 + \sum_{i=0}^{\infty} \sum_{j=s, i \neq j}^{\infty} \phi_1^i \phi_1^{j-s} 0 \\
&= \phi_1^s \left(\sum_{i=0}^{\infty} \phi_1^{2i} \sigma^2 \right) \\
&= \phi_1^s \gamma_0
\end{aligned}$$

Finally, the autocorrelations can be computed from ratios of autocovariances, $\rho_1 = \gamma_1 / \gamma_0 = \phi_1$ and $\rho_s = \gamma_s / \gamma_0 = \phi_1^s$.

4.A.3.2 MA(1)

The MA(1) model is the simplest non-degenerate time-series model considered in this course,

$$y_t = \theta_1 \varepsilon_{t-1} + \varepsilon_t$$

and the derivation of its autocorrelation function is trivial since there no backward substitution is required. The variance is

$$\begin{aligned}
\gamma_0 &= V[y_t] = E[y_t^2] \tag{4.136} \\
&= E[(\theta_1 \varepsilon_{t-1} + \varepsilon_t)^2] \\
&= E[\theta_1^2 \varepsilon_{t-1}^2 + 2\theta_1 \varepsilon_t \varepsilon_{t-1} + \varepsilon_t^2] \\
&= E[\theta_1^2 \varepsilon_{t-1}^2] + E[2\theta_1 \varepsilon_t \varepsilon_{t-1}] + E[\varepsilon_t^2] \\
&= \theta_1^2 \sigma^2 + 0 + \sigma^2 \\
&= \sigma^2(1 + \theta_1^2)
\end{aligned}$$

and the 1st autocovariance is

$$\begin{aligned}
\gamma_1 &= E[y_t y_{t-1}] \\
&= E[(\theta_1 \varepsilon_{t-1} + \varepsilon_t)(\theta_1 \varepsilon_{t-2} + \varepsilon_{t-1})] \\
&= E[\theta_1^2 \varepsilon_{t-1} \varepsilon_{t-2} + \theta_1 \varepsilon_{t-1}^2 + \theta_1 \varepsilon_t \varepsilon_{t-2} + \varepsilon_t \varepsilon_{t-1}] \\
&= E[\theta_1^2 \varepsilon_{t-1} \varepsilon_{t-2}] + E[\theta_1 \varepsilon_{t-1}^2] + E[\theta_1 \varepsilon_t \varepsilon_{t-2}] + E[\varepsilon_t \varepsilon_{t-1}] \\
&= 0 + \theta_1 \sigma^2 + 0 + 0 \\
&= \theta_1 \sigma^2
\end{aligned} \tag{4.137}$$

The 2nd(and higher) autocovariance is

$$\begin{aligned}
\gamma_2 &= E[y_t y_{t-2}] \\
&= E[(\theta_1 \varepsilon_{t-1} + \varepsilon_t)(\theta_1 \varepsilon_{t-3} + \varepsilon_{t-2})] \\
&= E[\theta_1^2 \varepsilon_{t-1} \varepsilon_{t-3} + \theta_1 \varepsilon_{t-1} \varepsilon_{t-2} + \theta_1 \varepsilon_t \varepsilon_{t-3} + \varepsilon_t \varepsilon_{t-2}] \\
&= E[\theta_1^2 \varepsilon_{t-1} \varepsilon_{t-3}] + E[\theta_1 \varepsilon_{t-1} \varepsilon_{t-2}] + E[\theta_1 \varepsilon_t \varepsilon_{t-3}] + E[\varepsilon_t \varepsilon_{t-2}] \\
&= 0 + 0 + 0 + 0 \\
&= 0
\end{aligned} \tag{4.138}$$

and the autocorrelations are $\rho_1 = \theta_1 / (1 + \theta_1^2)$, $\rho_s = 0$, $s \geq 2$.

4.A.3.3 ARMA(1,1)

An ARMA(1,1) process,

$$y_t = \phi_1 y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t$$

is stationary if $|\phi_1| < 1$ and $\{\varepsilon_t\}$ is white noise. The derivation of the variance and autocovariances is more tedious than for the AR(1) process. It should be noted that derivation is longer and more complex than solving the Yule-Walker equations.

Begin by computing the MA(∞) representation,

$$\begin{aligned}
y_t &= \phi_1 y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t \\
y_t &= \phi_1(\phi_1 y_{t-2} + \theta_1 \varepsilon_{t-2} + \varepsilon_{t-1}) + \theta_1 \varepsilon_{t-1} + \varepsilon_t \\
y_t &= \phi_1^2 y_{t-2} + \phi_1 \theta_1 \varepsilon_{t-2} + \phi_1 \varepsilon_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t \\
y_t &= \phi_1^2 (\phi_1 y_{t-3} + \theta_1 \varepsilon_{t-3} + \varepsilon_{t-2}) + \phi_1 \theta_1 \varepsilon_{t-2} + (\phi_1 + \theta_1) \varepsilon_{t-1} + \varepsilon_t \\
y_t &= \phi_1^3 y_{t-3} + \phi_1^2 \theta_1 \varepsilon_{t-3} + \phi_1^2 \varepsilon_{t-2} + \phi_1 \theta_1 \varepsilon_{t-2} + (\phi_1 + \theta_1) \varepsilon_{t-1} + \varepsilon_t \\
y_t &= \phi_1^3 (\phi_1 y_{t-4} + \theta_1 \varepsilon_{t-4} + \varepsilon_{t-3}) + \phi_1^2 \theta_1 \varepsilon_{t-3} + \phi_1 (\phi_1 + \theta_1) \varepsilon_{t-2} + (\phi_1 + \theta_1) \varepsilon_{t-1} + \varepsilon_t \\
y_t &= \phi_1^4 y_{t-4} + \phi_1^3 \theta_1 \varepsilon_{t-4} + \phi_1^3 \varepsilon_{t-3} + \phi_1^2 \theta_1 \varepsilon_{t-3} + \phi_1 (\phi_1 + \theta_1) \varepsilon_{t-2} + (\phi_1 + \theta_1) \varepsilon_{t-1} + \varepsilon_t \\
y_t &= \phi_1^4 y_{t-4} + \phi_1^3 \theta_1 \varepsilon_{t-4} + \phi_1^2 (\phi_1 + \theta_1) \varepsilon_{t-3} + \phi_1 (\phi_1 + \theta_1) \varepsilon_{t-2} + (\phi_1 + \theta_1) \varepsilon_{t-1} + \varepsilon_t \\
y_t &= \varepsilon_t + (\phi_1 + \theta_1) \varepsilon_{t-1} + \phi_1 (\phi_1 + \theta_1) \varepsilon_{t-2} + \phi_1^2 (\phi_1 + \theta_1) \varepsilon_{t-3} + \dots
\end{aligned} \tag{4.139}$$

$$y_t = \varepsilon_t + \sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \varepsilon_{t-1-i}$$

The primary issue is that the backward substitution form, unlike in the AR(1) case, is not completely symmetric. Specifically, ε_t has a different weight than the other shocks and does not follow the same pattern.

$$\begin{aligned}
\gamma_0 &= V[y_t] = E[y_t^2] \\
&= E \left[\left(\varepsilon_t + \sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \varepsilon_{t-1-i} \right)^2 \right] \\
&= E \left[(\varepsilon_t + (\phi_1 + \theta_1) \varepsilon_{t-1} + \phi_1 (\phi_1 + \theta_1) \varepsilon_{t-2} + \phi_1^2 (\phi_1 + \theta_1) \varepsilon_{t-3} + \dots)^2 \right] \\
&= E \left[\varepsilon_t^2 + 2\varepsilon_t \sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \varepsilon_{t-1-i} + \left(\sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \varepsilon_{t-1-i} \right)^2 \right] \\
&= E[\varepsilon_t^2] + E \left[2\varepsilon_t \sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \varepsilon_{t-1-i} \right] + E \left[\left(\sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \varepsilon_{t-1-i} \right)^2 \right] \\
&= \sigma^2 + 0 + E \left[\left(\sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \varepsilon_{t-1-i} \right)^2 \right] \\
&= \sigma^2 + E \left[\sum_{i=0}^{\infty} \phi_1^{2i} (\phi_1 + \theta_1)^2 \varepsilon_{t-1-i}^2 + \sum_{i=0}^{\infty} \sum_{j=0, j \neq i}^{\infty} \phi_1^i \phi_1^j (\phi_1 + \theta_1)^2 \varepsilon_{t-1-i} \varepsilon_{t-1-j} \right] \\
&= \sigma^2 + \sum_{i=0}^{\infty} \phi_1^{2i} (\phi_1 + \theta_1)^2 E[\varepsilon_{t-1-i}^2] + \sum_{i=0}^{\infty} \sum_{j=0, j \neq i}^{\infty} \phi_1^i \phi_1^j (\phi_1 + \theta_1)^2 E[\varepsilon_{t-1-i} \varepsilon_{t-1-j}] \\
&= \sigma^2 + \sum_{i=0}^{\infty} \phi_1^{2i} (\phi_1 + \theta_1)^2 \sigma^2 + \sum_{i=0}^{\infty} \sum_{j=0, j \neq i}^{\infty} \phi_1^i \phi_1^j (\phi_1 + \theta_1)^2 0 \\
&= \sigma^2 + \sum_{i=0}^{\infty} \phi_1^{2i} (\phi_1 + \theta_1)^2 \sigma^2 \\
&= \sigma^2 + \frac{(\phi_1 + \theta_1)^2 \sigma^2}{1 - \phi_1^2} \\
&= \sigma^2 \frac{1 - \phi_1^2 + (\phi_1 + \theta_1)^2}{1 - \phi_1^2} \\
&= \sigma^2 \frac{1 + \theta_1^2 + 2\phi_1 \theta_1}{1 - \phi_1^2}
\end{aligned} \tag{4.140}$$

The difficult step in this derivation is in aligning the ε_{t-i} since $\{\varepsilon_t\}$ is a white noise process. The autocovariance derivation is fairly involved (and presented in full detail).

$$\begin{aligned}
\gamma_1 &= E[y_t y_{t-1}] \tag{4.141} \\
&= E \left[\left(\varepsilon_t + \sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \varepsilon_{t-1-i} \right) \left(\varepsilon_{t-1} + \sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \varepsilon_{t-2-i} \right) \right] \\
&= E \left[(\varepsilon_t + (\phi_1 + \theta_1) \varepsilon_{t-1} + \phi_1 (\phi_1 + \theta_1) \varepsilon_{t-2} + \phi_1^2 (\phi_1 + \theta_1) \varepsilon_{t-3} + \dots) \times \right. \\
&\quad \left. (\varepsilon_{t-1} + (\phi_1 + \theta_1) \varepsilon_{t-2} + \phi_1 (\phi_1 + \theta_1) \varepsilon_{t-3} + \phi_1^2 (\phi_1 + \theta_1) \varepsilon_{t-4} + \dots) \right] \\
&= E \left[\varepsilon_t \varepsilon_{t-1} + \sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \varepsilon_t \varepsilon_{t-2-i} + \sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \varepsilon_{t-1} \varepsilon_{t-1-i} \right. \\
&\quad \left. + \left(\sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \varepsilon_{t-1-i} \right) \left(\sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \varepsilon_{t-2-i} \right) \right] \\
&= E[\varepsilon_t \varepsilon_{t-1}] + E \left[\sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \varepsilon_t \varepsilon_{t-2-i} \right] + E \left[\sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \varepsilon_{t-1} \varepsilon_{t-1-i} \right] \\
&\quad + E \left[\left(\sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \varepsilon_{t-1-i} \right) \left(\sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \varepsilon_{t-2-i} \right) \right] \\
&= 0 + 0 + (\phi_1 + \theta_1) \sigma^2 + E \left[\left(\sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \varepsilon_{t-1-i} \right) \left(\sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \varepsilon_{t-2-i} \right) \right] \\
&= (\phi_1 + \theta_1) \sigma^2 + E \left[\sum_{i=0}^{\infty} \phi_1^{2i+1} (\phi_1 + \theta_1)^2 \varepsilon_{t-2-i}^2 + \sum_{i=0}^{\infty} \sum_{j=0, i \neq j+1}^{\infty} \phi_1^i \phi_1^j (\phi_1 + \theta_1)^2 \varepsilon_{t-1-i} \varepsilon_{t-2-i} \right] \\
&= (\phi_1 + \theta_1) \sigma^2 + E \left[\sum_{i=0}^{\infty} \phi_1^{2i+1} (\phi_1 + \theta_1)^2 \varepsilon_{t-2-i}^2 \right] + E \left[\sum_{i=0}^{\infty} \sum_{j=0, i \neq j+1}^{\infty} \phi_1^i \phi_1^j (\phi_1 + \theta_1)^2 \varepsilon_{t-1-i} \varepsilon_{t-2-i} \right] \\
&= (\phi_1 + \theta_1) \sigma^2 + E \left[\phi_1 \sum_{i=0}^{\infty} \phi_1^{2i} (\phi_1 + \theta_1)^2 \varepsilon_{t-2-i}^2 \right] + 0 \\
&= (\phi_1 + \theta_1) \sigma^2 + \phi_1 \sum_{i=0}^{\infty} \phi_1^{2i} (\phi_1 + \theta_1)^2 E[\varepsilon_{t-2-i}^2] \\
&= (\phi_1 + \theta_1) \sigma^2 + \phi_1 \sum_{i=0}^{\infty} \phi_1^{2i} (\phi_1 + \theta_1)^2 \sigma^2 \\
&= (\phi_1 + \theta_1) \sigma^2 + \phi_1 \frac{(\phi_1 + \theta_1)^2 \sigma^2}{1 - \phi_1^2} \\
&= \frac{\sigma^2 \left[(1 - \phi_1^2) (\phi_1 + \theta_1) + \phi_1 (\phi_1 + \theta_1)^2 \right]}{1 - \phi_1^2} \\
&= \frac{\sigma^2 (\phi_1 + \theta_1 - \phi_1^3 - \phi_1^2 \theta_1 + \phi_1^3 + 2\phi_1^2 \theta_1 - \phi_1 \theta_1^2)}{1 - \phi_1^2}
\end{aligned}$$

$$\begin{aligned}
&= \frac{\sigma^2 [\phi_1 + \theta_1 + \phi_1^2 \theta_1 - \phi_1 \theta_1^2]}{1 - \phi_1^2} \\
&= \frac{\sigma^2 (\phi_1 + \theta_1) (\phi_1 \theta_1 + 1)}{1 - \phi_1^2}
\end{aligned}$$

The most difficult step in this derivation is in showing that $E[\sum_{i=0}^{\infty} \phi_1^i (\phi_1 + \theta_1) \varepsilon_{t-1-i}] = \sigma^2 (\phi_1 + \theta_1)$ since there is one ε_{t-1-i} which is aligned to ε_{t-1} (i.e. when $i = 0$), and so the autocorrelations may be derived,

$$\begin{aligned}
\rho_1 &= \frac{\frac{\sigma^2 (\phi_1 + \theta_1) (\phi_1 \theta_1 + 1)}{1 - \phi_1^2}}{\frac{\sigma^2 (1 + \theta_1^2 + 2\phi_1 \theta_1)}{1 - \phi_1^2}} \\
&= \frac{(\phi_1 + \theta_1) (\phi_1 \theta_1 + 1)}{(1 + \theta_1^2 + 2\phi_1 \theta_1)}
\end{aligned} \tag{4.142}$$

and the remaining autocorrelations can be computed using the recursion, $\rho_s = \phi_1 \rho_{s-1}$, $s \geq 2$.

Shorter Problems

Problem 4.1. What is the optimal 3-step forecast from the ARMA(1,2), $y_t = \phi_0 + \phi_1 y_{t-1} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \varepsilon_t$, where ε_t is a mean 0 white noise process?

Problem 4.2. What are the expected values for α , β and γ when a forecasting model is well specified in the Mincer-Zarnowitz regression,

$$y_{t+h} = \alpha + \beta \hat{y}_{t+h|t} + \gamma x_t + \eta_{t+h}.$$

Provide an explanation for why these values should be expected.

Problem 4.3. What are the consequences of using White or Newey-West to estimate the covariance in a linear regression when the errors are serially uncorrelated and homoskedastic?

Problem 4.4. What are the 1-step and 2-step optimal forecasts for the conditional mean when $y_t = \phi_0 + \phi_1 y_{t-1} + \varepsilon_t$ where $\varepsilon_t \stackrel{i.i.d.}{\sim} N(0, 1)$?

Problem 4.5. Is the sum of two white noise processes, $\varepsilon_t = \eta_t + \nu_t$ necessarily a white noise process?

Problem 4.6. What are the 1-step and 2-step optimal mean square forecast errors when $y_t = \phi_0 + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \varepsilon_t$ where $\varepsilon_t \stackrel{i.i.d.}{\sim} N(0, 1)$?

Problem 4.7. Outline the steps needed to perform a Diebold-Mariano test that two models for the conditional mean are equivalent (in the MSE sense).

Problem 4.8. Justify a reasonable model for each of these time series in Figure 4.15 using information in the autocorrelation and partial autocorrelation plots. In each set of plots, the left most panel shows that data ($T = 100$). The middle panel shows the sample autocorrelation with 95% confidence bands. The right panel shows the sample partial autocorrelation for the data with 95% confidence bands.

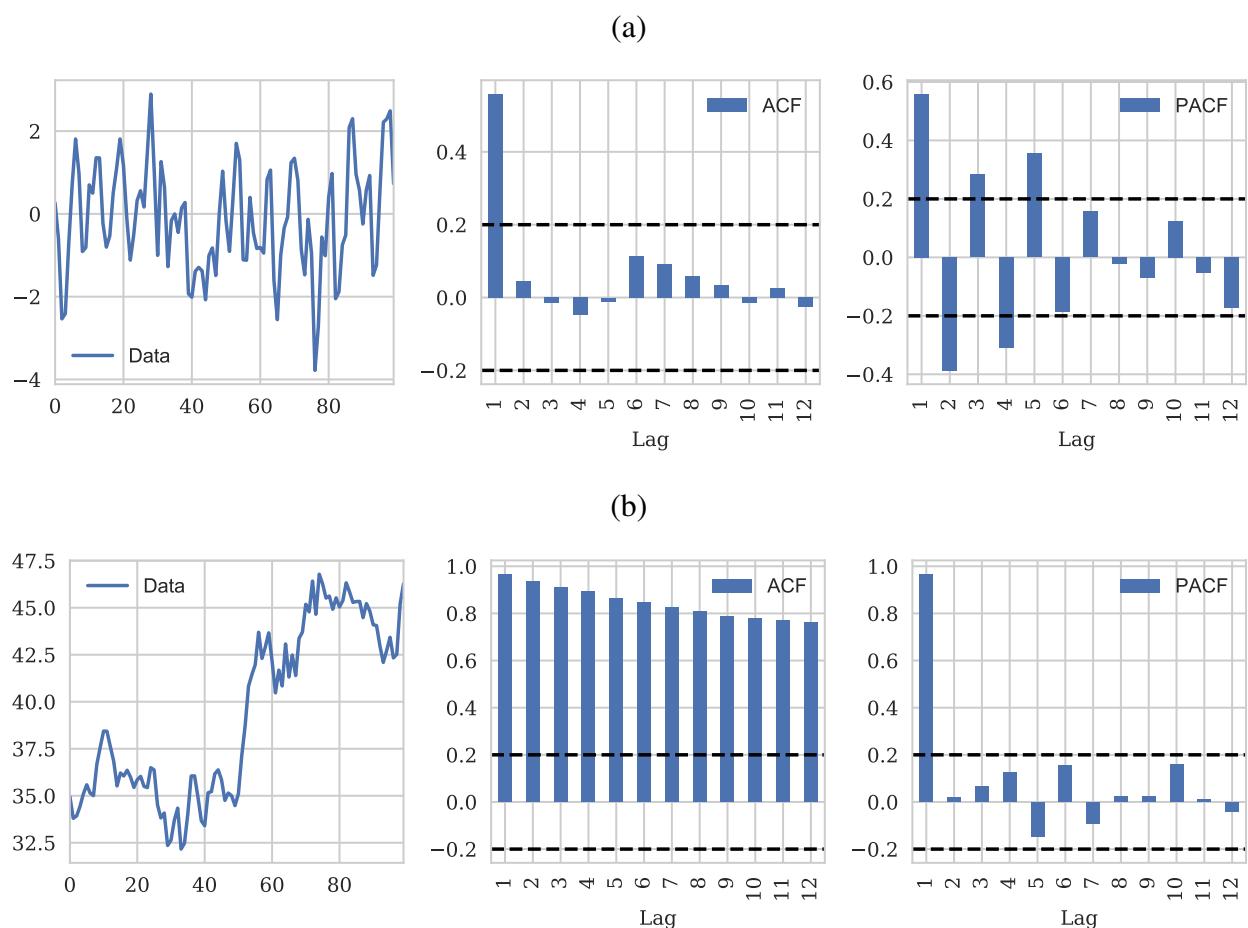


Figure 4.15: Plots for question 2(b).

Longer Exercises

Exercise 4.1. Answer the following questions:

1. Under what conditions on the parameters and errors are the following processes covariance stationary?
 - (a) $y_t = \phi_0 + \varepsilon_t$
 - (b) $y_t = \phi_0 + \phi_1 y_{t-1} + \varepsilon_t$
 - (c) $y_t = \phi_0 + \theta_1 \varepsilon_{t-1} + \varepsilon_t$
 - (d) $y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$
 - (e) $y_t = \phi_0 + \phi_2 y_{t-2} + \varepsilon_t$
 - (f) $y_t = \phi_0 + \phi_1 y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t$
2. Is the sum of two white noise processes, $v_t = \varepsilon_t + \eta_t$, necessarily a white noise process? If so, verify that the properties of a white noise are satisfied. If not, show why and describe any further assumptions required for the sum to be a white noise process.

Exercise 4.2. Consider an AR(1)

$$y_t = \phi_0 + \phi_1 y_{t-1} + \varepsilon_t$$

1. What is a minimal set of assumptions sufficient to ensure $\{y_t\}$ is covariance stationary if $\{\varepsilon_t\}$ is an i.i.d. sequence?
2. What are the values of the following quantities?
 - (a) $E[y_{t+1}]$
 - (b) $E_t[y_{t+1}]$
 - (c) $V[y_{t+1}]$
 - (d) $V_t[y_{t+1}]$
 - (e) ρ_{-1}
 - (f) ρ_2

Exercise 4.3. Consider an MA(1)

$$y_t = \phi_0 + \theta_1 \varepsilon_{t-1} + \varepsilon_t$$

1. What is a minimal set of assumptions sufficient to ensure $\{y_t\}$ is covariance stationary if $\{\varepsilon_t\}$ is an i.i.d. sequence?
2. What are the values of the following quantities?
 - (a) $E[y_{t+1}]$
 - (b) $E_t[y_{t+1}]$
 - (c) $V[y_{t+1}]$

- (d) $V_t[y_{t+1}]$
 (e) ρ_{-1}
 (f) ρ_2
3. Suppose you were trying to differentiate between an AR(1) and an MA(1) but could not estimate any regressions. What would you do?

Exercise 4.4. Consider an MA(2)

$$y_t = \mu + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \varepsilon_t$$

1. What is a minimal set of assumptions sufficient to ensure $\{y_t\}$ is covariance stationary if $\{\varepsilon_t\}$ is an i.i.d. sequence?
2. What are the values of the following quantities?
 - (a) $E[y_{t+1}]$
 - (b) $E_t[y_{t+1}]$
 - (c) $V[y_{t+1}]$
 - (d) $V_t[y_{t+1}]$
 - (e) ρ_{-1}
 - (f) ρ_2
 - (g) ρ_3

Exercise 4.5. Answer the following questions:

1. For each of the following processes, find $E_t[y_{t+1}]$. You can assume $\{\varepsilon_t\}$ is a mean zero i.i.d. sequence.
 - (a) $y_t = \phi_0 + \phi_1 y_{t-1} + \varepsilon_t$
 - (b) $y_t = \phi_0 + \theta_1 \varepsilon_{t-1} + \varepsilon_t$
 - (c) $y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$
 - (d) $y_t = \phi_0 + \phi_2 y_{t-2} + \varepsilon_t$
 - (e) $y_t = \phi_0 + \phi_1 y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t$
2. For (a), (c) and (e), derive the h -step ahead forecast, $E_t[y_{t+h}]$. What is the long run behavior of the forecast in each case?
3. The forecast error variance is defined as $E[(y_{t+h} - E_t[y_{t+h}])^2]$. Find an explicit expression for the forecast error variance for (a) and (c).

Exercise 4.6. Answer the following questions:

1. What are the characteristic equations for the above systems?

- (a) $y_t = 1 + .6y_{t-1} + x_t$
 (b) $y_t = 1 + .8y_{t-2} + x_t$
 (c) $y_t = 1 + .6y_{t-1} + .3y_{t-2} + x_t$
 (d) $y_t = 1 + 1.2y_{t-1} + .2y_{t-2} + x_t$
 (e) $y_t = 1 + 1.4y_{t-1} + .24y_{t-2} + x_t$
 (f) $y_t = 1 - .8y_{t-1} + .2y_{t-2} + x_t$
2. Compute the roots for the characteristic equation? Which are convergent? Which are explosive? Are any stable or metastable?

Exercise 4.7. Suppose that y_t follows a random walk then $\Delta y_t = y_t - y_{t-1}$ is stationary.

1. Is $y_t - y_{t-j}$ for and $j \geq 2$ stationary?
 2. If it is and $\{\varepsilon_t\}$ is an i.i.d. sequence of standard normals, what is the distribution of $y_t - y_{t-j}$?
 3. What is the joint distribution of $y_t - y_{t-j}$ and $y_{t-h} - y_{t-j-h}$ (Note: The derivation for an arbitrary h is challenging)?
- Note:** If it helps in this problem, consider the case where $j = 2$ and $h = 1$.

Exercise 4.8. Outline the steps needed to perform a unit root test on as time-series of FX rates. Be sure to detail the any important considerations that may affect the test.

Exercise 4.9. Answer the following questions:

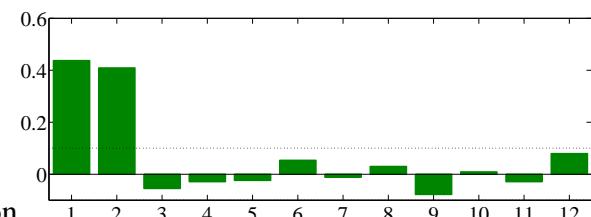
1. How are the autocorrelations and partial autocorrelations useful in building a model?
2. Suppose you observe the three sets of ACF/PACF in figure 4.16. What ARMA specification would you expect in each case. Note: Dashed line indicates the 95% confidence interval for a test that the autocorrelation or partial autocorrelation is 0.
3. Describe the three methods of model selection discussed in class: general-to-specific, specific-to-general and the use of information criteria (Schwarz/Bayesian Information Criteria and/or Akaike Information Criteria). When might each be preferred to the others?
4. Describe the Wald, Lagrange Multiplier (Score) and Likelihood ratio tests. What aspect of a model does each test? What are the strengths and weaknesses of each?

Exercise 4.10. Answer the following questions about forecast errors.

1. Let $y_t = \phi_0 + \phi_1 y_{t-1} + \varepsilon_t$ with the usual assumptions on $\{\varepsilon_t\}$. Derive an explicit expression for the 1-step and 2-step ahead forecast errors, $e_{t+h|t} = y_{t+h} - \hat{y}_{t+h|t}$ where $\hat{y}_{t+h|t}$ is the MSE optimal forecast where $h = 1$ or $h = 2$ (what is the MSE optimal forecast?).
2. What is the autocorrelation function of a time-series of forecast errors $\{e_{t+h|t}\}$, $h = 1$ or $h = 2$. (Hint: Use the formula you derived above)

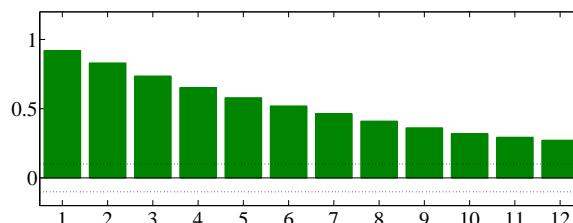
**Autocorrelation and Partial Autocorrelation function
ACF**

(a)



The middle panel shows the sample autocorrelation

(b)



(c)

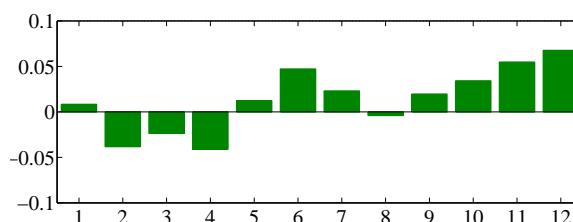


Figure 4.16: The ACF and PACF of three stochastic processes. Use these to answer question 4.9.

3. Can you generalize the above to a generic h ? (In other words, leave the solution as a function of h).
4. How could you test whether the forecast has excess dependence using an ARMA model?

Exercise 4.11. Answer the following questions.

1. Outline the steps needed to determine whether a time series $\{y_t\}$ contains a unit root. Be certain to discuss the important considerations at each step, if any.
2. If y_t follows a pure random walk driven by white noise innovations then $\Delta y_t = y_t - y_{t-1}$ is stationary.
 - (a) Is $y_t - y_{t-j}$ for and $j \geq 2$ stationary?
 - (b) If it is and $\{\varepsilon_t\}$ is an i.i.d. sequence of standard normals, what is the distribution of $y_t - y_{t-j}$?
 - (c) What is the joint distribution of $y_t - y_{t-j}$ and $y_{t-h} - y_{t-j-h}$?
3. Let $y_t = \phi_0 + \phi_1 y_{t-1} + \varepsilon_t$ where $\{\varepsilon_t\}$ is a WN process.
 - (a) Derive an explicit expression for the 1-step and 2-step ahead forecast errors, $e_{t+h|t} = y_{t+h} - \hat{y}_{t+h|t}$ where $\hat{y}_{t+h|t}$ is the MSE optimal forecast where $h = 1$ or $h = 2$.
 - (b) What is the autocorrelation function of a time-series of forecast errors $\{e_{t+h|t}\}$ for $h = 1$ and $h = 2$?
 - (c) Generalize the above to a generic h ? (In other words, leave the solution as a function of h).
 - (d) How could you test whether the forecast has excess dependence using an ARMA model?

Exercise 4.12. Suppose

$$y_t = \phi_0 + \phi_1 y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t$$

where $\{\varepsilon_t\}$ is a white noise process.

1. Precisely describe the two types of stationarity.
2. Why is stationarity a useful property?
3. What conditions on the model parameters are needed for $\{y_t\}$ to be covariance stationary?
4. Describe the Box-Jenkins methodology for model selection.
Now suppose that $\phi_1 = 1$ and that ε_t is homoskedastic.
 5. What is $E_t[y_{t+1}]$?
 6. What is $E_t[y_{t+2}]$?
 7. What can you say about $E_t[y_{t+h}]$ for $h > 2$?

8. What is $V_t[y_{t+1}]$?
9. What is $V_t[y_{t+2}]$?
10. What is the first autocorrelation, ρ_1 ?

Exercise 4.13. Which of the following models are covariance stationary, assuming $\{\varepsilon_t\}$ is a mean-zero white noise process. If the answer is conditional, explain the conditions required. In any case, explain your answer:

1. $y_t = \phi_0 + 0.8y_{t-1} + 0.2y_{t-2} + \varepsilon_t$
2. $y_t = \phi_0 + \phi_1 I_{[t>200]} + \varepsilon_t$
3. $y_t = \alpha t + 0.8\varepsilon_{t-1} + \varepsilon_t$
4. $y_t = 4\varepsilon_{t-1} + 9\varepsilon_{t-2} + \varepsilon_t$
5. $y_t = \varepsilon_t + \sum_{j=1}^{\infty} \gamma_j \varepsilon_{t-j}$

Exercise 4.14. Answer the following questions:

1. Consider the AR(2)

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$$

- (a) Rewrite the model with Δy_t on the left-hand side and y_{t-1} and Δy_{t-1} on the right-hand side.
- (b) What restrictions are needed on ϕ_1 and ϕ_2 for this model to collapse to an AR(1) in the first differences?
- (c) When the model collapses, what does this tell you about y_t ?

2. Discuss the important issues when testing for unit roots in economic time-series.

Exercise 4.15. In which of the following models are the $\{y_t\}$ covariance stationary, assuming $\{\varepsilon_t\}$ is a mean-zero white noise process. If the answer is conditional, explain the conditions required. In any case, explain your answer:

1. $\Delta y_t = -0.2y_{t-1} + \varepsilon_t$
2. $y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$
3. $y_t = \phi_0 + 0.1x_{t-1} + \varepsilon_t, x_t = x_{t-1} + \varepsilon_t$
4. $y_t = 0.8y_{t-1} + \varepsilon_t$

Exercise 4.16. Suppose

$$y_t = \phi_0 + \phi_1 y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t$$

where $\{\varepsilon_t\}$ is a white noise process.

1. Precisely describe the two types of stationarity.
2. Why is stationarity a useful property?
3. What conditions on the model parameters are needed for $\{y_t\}$ to be covariance stationary?
4. Describe the Box-Jenkins methodology for model selection.
5. Now suppose that $\phi_1 = 1$ and that ε_t is homoskedastic.
6. What is $E_t[y_{t+1}]$?
7. What is $E_t[y_{t+2}]$?
8. What can you say about $E_t[y_{t+h}]$ for $h > 2$?
9. What is $V_t[y_{t+1}]$?
10. What is $V_t[y_{t+2}]$?
11. What is the first autocorrelation, ρ_1 ?

Exercise 4.17. Answer the following questions.

1. Suppose $y_t = \phi_0 + \phi_1 y + \phi_2 y_{t-2} + \varepsilon_t$ where $\{\varepsilon_t\}$ is a white noise process.
2. Write this model in companion form.
 - (a) Using the companion form, derive expressions for the first two autocovariances of y_t . (It is not necessary to explicitly solve them in scalar form)
 - (b) Using the companion form, determine the formal conditions for ϕ_1 and ϕ_2 to for $\{y_t\}$ to be covariance stationary. You can use the result that when A is a 2 by 2 matrix, its eigenvalues solve the two equations

$$\begin{aligned}\lambda_1 \lambda_2 &= a_{11}a_{22} - a_{12}a_{21} \\ \lambda_1 + \lambda_2 &= a_{11} + a_{22}\end{aligned}$$

Exercise 4.18. Justify a reasonable model for each of these time series in Figure 4.17 using information in the autocorrelation and partial autocorrelation plots. In each set of plots, the left most panel shows that data ($T = 100$). The middle panel shows the sample autocorrelation with 95% confidence bands. The right panel shows the sample partial autocorrelation for the data with 95% confidence bands.

1. Panel (a)
2. Panel (b)
3. Panel (c)

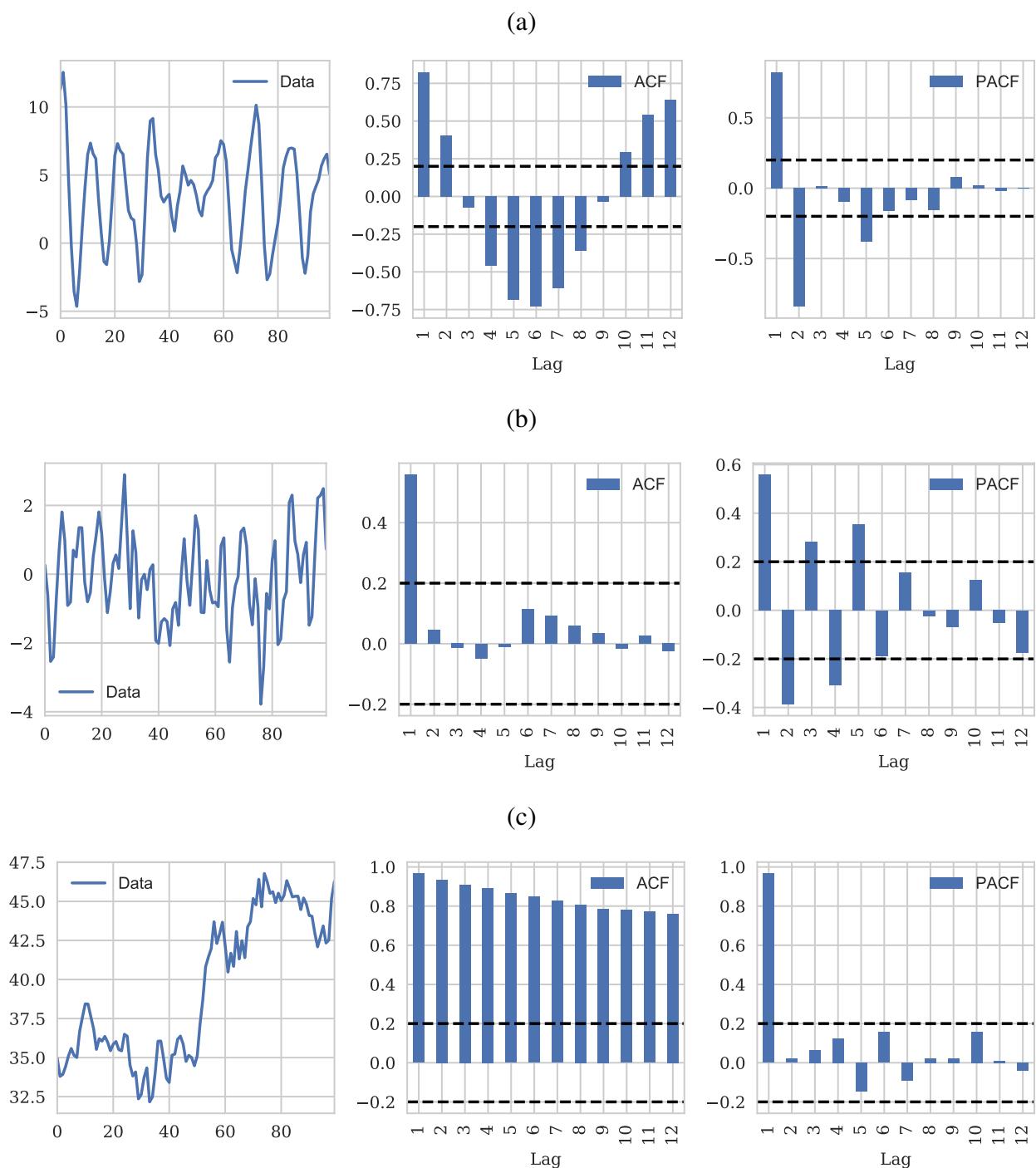


Figure 4.17: Plots for question 2(b).

Chapter 5

Analysis of Multiple Time Series

The alternative reference for the material in this chapter is Enders (2004) (chapters 5 and 6). Chapters 10-11 and 18-19 in Hamilton (1994) provide a more technical treatment of the material.

Multivariate time-series analysis extends many of the ideas of univariate time-series analysis to systems of equations. The primary model used in multivariate time-series analysis is the vector autoregression (VAR). Many properties of autoregressive processes extend naturally to multivariate time-series using a slight change in notation and results from linear algebra. This chapter examines the properties of vector time-series models, estimation and identification and introduces two new concepts: Granger Causality and the Impulse Response Function. The chapter concludes by examining models of contemporaneous relationships between two or more time-series in the framework of cointegration, spurious regression and cross-sectional regression of stationary time-series.

In many applications, analyzing a time-series in isolation is a reasonable choice; in others, univariate analysis is insufficient to capture the complex dynamics among interrelated time series. For example, Campbell (1996) links financially interesting variables, including stock returns and the default premium, in a multivariate system where shocks to one variable propagate to the others. The vector autoregression (VAR) is the standard model used to model multiple *stationary* time-series. If the time series are not stationary, a different type of analysis, cointegration, is used.

5.1 Vector Autoregressions

Vector autoregressions are remarkably similar to univariate autoregressions, and most results carry over by replacing scalars with matrices and scalar operations with their linear algebra equivalent.

5.1.1 Definition

The definition of a vector autoregression is nearly identical to that of a univariate autoregression.

Definition 5.1 (Vector Autoregression of Order P). A P^{th} order vector autoregression, written $\text{VAR}(P)$, is a process that evolves according to

$$\mathbf{y}_t = \Phi_0 + \Phi_1 \mathbf{y}_{t-1} + \Phi_2 \mathbf{y}_{t-2} + \dots + \Phi_P \mathbf{y}_{t-P} + \boldsymbol{\varepsilon}_t \quad (5.1)$$

where \mathbf{y}_t is a k by 1 vector stochastic process, Φ_0 is a k by 1 vector of intercept parameters, Φ_j , $j = 1, \dots, P$ are k by k parameter matrices and ε_t is a vector white noise process with the additional assumption that $E_{t-1}[\varepsilon_t] = \mathbf{0}$.

A VAR(P) reduces to an AR(P) when $k = 1$ so that \mathbf{y}_t and the coefficient matrices, Φ_j , are scalars. A vector white noise process extends the three properties of a univariate white noise process to a vector; it is mean zero, has finite covariance and is uncorrelated with its past. The components of a vector white noise process are not assumed to be *contemporaneously* uncorrelated.

Definition 5.2 (Vector White Noise Process). A k by 1 vector-valued stochastic process, $\{\varepsilon_t\}$ is a vector white noise if

$$\begin{aligned} E[\varepsilon_t] &= \mathbf{0}_k \\ E[\varepsilon_t \varepsilon'_{t-s}] &= \mathbf{0}_{k \times k} \\ E[\varepsilon_t \varepsilon'_t] &= \Sigma \end{aligned} \tag{5.2}$$

for all t where Σ is a finite positive definite matrix.

The simplest VAR is a first-order bivariate specification which is equivalently expressed as

$$\mathbf{y}_t = \Phi_0 + \Phi_1 \mathbf{y}_{t-1} + \varepsilon_t,$$

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} \phi_{1,0} \\ y_{2,0} \end{bmatrix} + \begin{bmatrix} \phi_{11,1} & \phi_{12,1} \\ \phi_{21,1} & \phi_{22,1} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix},$$

or

$$\begin{aligned} y_{1,t} &= \phi_{1,0} + \phi_{11,1}y_{1,t-1} + \phi_{12,1}y_{2,t-1} + \varepsilon_{1,t} \\ y_{2,t} &= \phi_{2,0} + \phi_{21,1}y_{1,t-1} + \phi_{22,1}y_{2,t-1} + \varepsilon_{2,t}. \end{aligned}$$

Each element of \mathbf{y}_t is a function of each element of \mathbf{y}_{t-1} .

5.1.2 Properties of a VAR(1)

The properties of the VAR(1) are straightforward to derive. Importantly, section 5.2 shows that all VAR(P)s can be rewritten as a VAR(1), and so the properties of any VAR follow from those of a first-order VAR.

5.1.2.1 Stationarity

A VAR(1), driven by vector white noise shocks,

$$\mathbf{y}_t = \Phi_0 + \Phi_1 \mathbf{y}_{t-1} + \varepsilon_t$$

is covariance stationary if the eigenvalues of Φ_1 are less than 1 in modulus.¹ In the univariate case, this statement is equivalent to the condition $|\phi_1| < 1$. Assuming the eigenvalues of Φ_1 are less than one in absolute value, backward substitution can be used to show that

¹The definition of an eigenvalue is:

$$\mathbf{y}_t = \sum_{i=0}^{\infty} \Phi_1^i \Phi_0 + \sum_{i=0}^{\infty} \Phi_1^i \boldsymbol{\varepsilon}_{t-i} \quad (5.3)$$

which, applying Theorem 5.3, is equivalent to

$$\mathbf{y}_t = (\mathbf{I}_k - \Phi_1)^{-1} \Phi_0 + \sum_{i=0}^{\infty} \Phi_1^i \boldsymbol{\varepsilon}_{t-i} \quad (5.4)$$

where the eigenvalue condition ensures that Φ_1^i converges to zero as i grows large.

5.1.2.2 Mean

Taking expectations of \mathbf{y}_t expressed in the backward substitution form yields

$$\begin{aligned} E[\mathbf{y}_t] &= E\left[(\mathbf{I}_k - \Phi_1)^{-1} \Phi_0\right] + E\left[\sum_{i=0}^{\infty} \Phi_1^i \boldsymbol{\varepsilon}_{t-i}\right] \\ &= (\mathbf{I}_k - \Phi_1)^{-1} \Phi_0 + \sum_{i=0}^{\infty} \Phi_1^i E[\boldsymbol{\varepsilon}_{t-i}] \\ &= (\mathbf{I}_k - \Phi_1)^{-1} \Phi_0 + \sum_{i=0}^{\infty} \Phi_1^i \mathbf{0} \\ &= (\mathbf{I}_k - \Phi_1)^{-1} \Phi_0 \end{aligned} \quad (5.5)$$

The mean of a VAR process resembles that of a univariate AR(1), $(1 - \phi_1)^{-1} \phi_0$.² The long-run mean depends on the intercept, Φ_0 , and the inverse of Φ_1 . The magnitude of the inverse is determined by

Definition 5.3 (Eigenvalue). λ is an eigenvalue of a square matrix \mathbf{A} if and only if $|\mathbf{A} - \lambda \mathbf{I}_n| = 0$ where $|\cdot|$ denotes determinant.

Definition 5.4. Eigenvalues play a unique role in the matrix power operator.

Theorem 5.1 (Singular Value Decomposition). *Let \mathbf{A} be an n by n real-valued matrix. Then \mathbf{A} can be decomposed as $\mathbf{A} = \mathbf{U}\Lambda\mathbf{V}'$ where $\mathbf{V}'\mathbf{U} = \mathbf{U}'\mathbf{V} = \mathbf{I}_n$ and Λ is a diagonal matrix containing the eigenvalues of \mathbf{A} .*

Theorem 5.2 (Matrix Power). *Let \mathbf{A} be an n by n real-valued matrix. Then $\mathbf{A}^m = \mathbf{A}\mathbf{A}\dots\mathbf{A} = \mathbf{U}\Lambda\mathbf{V}'\mathbf{U}\Lambda\mathbf{V}'\dots\mathbf{U}\Lambda\mathbf{V}' = \mathbf{U}\Lambda^m\mathbf{V}'$ where Λ^m is a diagonal matrix containing each eigenvalue of \mathbf{A} raised to the power m .*

The essential properties of eigenvalues for applications to VARs are given in the following theorem:

Theorem 5.3 (Convergent Matrices). *Let \mathbf{A} be an n by n matrix. Then the following statements are equivalent*

- $\mathbf{A}^m \rightarrow 0$ as $m \rightarrow \infty$.
- All eigenvalues of \mathbf{A} , λ_i , $i = 1, 2, \dots, n$, are less than 1 in modulus ($|\lambda_i| < 1$).
- The series $\sum_{i=0}^m \mathbf{A}^m = \mathbf{I}_n + \mathbf{A} + \mathbf{A}^2 + \dots + \mathbf{A}^m \rightarrow (\mathbf{I}_n - \mathbf{A})^{-1}$ as $m \rightarrow \infty$.

Note: Replacing \mathbf{A} with a scalar a produces many familiar results: $a^m \rightarrow 0$ as $m \rightarrow \infty$ (property 1) and $\sum_{i=0}^m a^i \rightarrow (1-a)^{-1}$ as $m \rightarrow \infty$ (property 3) as long as $|a| < 1$ (property 2).

²When a is a scalar where $|a| < 1$, then $\sum_{i=0}^{\infty} a^i = 1/(1-a)$. This result extends to a $k \times k$ square matrix \mathbf{A} when all of the eigenvalues of \mathbf{A} are less than 1, so that $\sum_{i=0}^{\infty} \mathbf{A}^i = (\mathbf{I}_k - \mathbf{A})^{-1}$.

the eigenvalues of Φ_1 , and if any eigenvalue is close to one, then $(\mathbf{I}_k - \Phi_1)^{-1}$ is large in magnitude and, all things equal, the unconditional mean is larger. Similarly, if $\Phi_1 = \mathbf{0}$, then the mean is Φ_0 since $\{\mathbf{y}_t\}$ is a constant plus white noise.

5.1.2.3 Variance

Before deriving the variance of a VAR(1), it is useful to express a VAR in *deviations* form. Define $\mu = E[\mathbf{y}_t]$ to be the unconditional expectation (assumed it is finite). The deviations form of the VAR(P)

$$\mathbf{y}_t = \Phi_0 + \Phi_1 \mathbf{y}_{t-1} + \Phi_2 \mathbf{y}_{t-2} + \dots + \Phi_P \mathbf{y}_{t-P} + \varepsilon_t$$

is

$$\begin{aligned}\mathbf{y}_t - \mu &= \Phi_1 (\mathbf{y}_{t-1} - \mu) + \Phi_2 (\mathbf{y}_{t-2} - \mu) + \dots + \Phi_P (\mathbf{y}_{t-P} - \mu) + \varepsilon_t \\ \tilde{\mathbf{y}}_t &= \Phi_1 \tilde{\mathbf{y}}_{t-1} + \Phi_2 \tilde{\mathbf{y}}_{t-2} + \dots + \Phi_P \tilde{\mathbf{y}}_{t-P} + \varepsilon_t.\end{aligned}\quad (5.6)$$

The deviations form is mean $\mathbf{0}$ by construction, and so the backward substitution form in a VAR(1) is

$$\tilde{\mathbf{y}}_t = \sum_{i=1}^{\infty} \Phi_1^i \varepsilon_{t-i}. \quad (5.7)$$

The deviations form translates the VAR from its original mean, μ , to a mean of $\mathbf{0}$. The process written in deviations form has the same dynamics and shocks, and so can be used to derive the long-run covariance and autocovariances and to simplify multistep forecasting. The long-run covariance is derived using the backward substitution form so that

$$\begin{aligned}E[(\mathbf{y}_t - \mu)(\mathbf{y}_t - \mu)'] &= E[\tilde{\mathbf{y}}_t \tilde{\mathbf{y}}_t'] = E\left[\left(\sum_{i=0}^{\infty} \Phi_1^i \varepsilon_{t-i}\right)\left(\sum_{i=0}^{\infty} \Phi_1^i \varepsilon_{t-i}\right)'\right] \\ &= E\left[\sum_{i=0}^{\infty} \Phi_1^i \varepsilon_{t-i} \varepsilon'_{t-i} (\Phi_1')'\right] \quad (\text{Since } \varepsilon_t \text{ is WN}) \\ &= \sum_{i=0}^{\infty} \Phi_1^i E[\varepsilon_{t-i} \varepsilon'_{t-i}] (\Phi_1')' \\ &= \sum_{i=0}^{\infty} \Phi_1^i \Sigma (\Phi_1')' \\ \text{vec}(E[(\mathbf{y}_t - \mu)(\mathbf{y}_t - \mu)']) &= (\mathbf{I}_{k^2} - \Phi_1 \otimes \Phi_1)^{-1} \text{vec}(\Sigma)\end{aligned}\quad (5.8)$$

where $\mu = (\mathbf{I}_k - \Phi_1)^{-1} \Phi_0$. The similarity between the long-run covariance of a VAR(1) and the long-run variance of a univariate autoregression, $\sigma^2 / (1 - \phi_1^2)$, are less pronounced. The difference between these expressions arises since matrix multiplication is non-commutative ($\mathbf{AB} \neq \mathbf{BA}$, in general). The

final line makes use of the *vec* (vector) operator to compactly express the long-run covariance. The *vec* operator and a Kronecker product stack the elements of a matrix product into a single column.³ The eigenvalues of Φ_1 also affect the long-run covariance, and if any are close to 1, the long-run covariance is large since the maximum eigenvalue determines the persistence of shocks. All things equal, more persistence lead to larger long-run covariances since the effect of any shock last longer.

5.1.2.4 Autocovariance

The autocovariances of a vector-valued stochastic process are defined

Definition 5.7 (Autocovariance). The autocovariance matrices of k by 1 vector-valued covariance stationary stochastic process $\{\mathbf{y}_t\}$ are defined

$$\Gamma_s = E[(\mathbf{y}_t - \mu)(\mathbf{y}_{t-s} - \mu)'] \quad (5.10)$$

and

$$\Gamma_{-s} = E[(\mathbf{y}_t - \mu)(\mathbf{y}_{t+s} - \mu)'] \quad (5.11)$$

where $\mu = E[\mathbf{y}_t] = E[\mathbf{y}_{t-j}] = E[\mathbf{y}_{t+j}]$.

The structure of the autocovariance function is the first significant deviation from the univariate time-series analysis in chapter 4. Vector autocovarianes are reflected, and so are symmetric only when transposed. Specifically,

³The *vec* of a matrix \mathbf{A} is defined:

Definition 5.5 (*vec*). Let $\mathbf{A} = [a_{ij}]$ be an m by n matrix. The *vec* operator (also known as the *stack* operator) is defined

$$\text{vec } \mathbf{A} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{bmatrix} \quad (5.9)$$

where \mathbf{a}_j is the j^{th} column of the matrix \mathbf{A} .

The Kronecker Product is defined:

Definition 5.6 (Kronecker Product). Let $\mathbf{A} = [a_{ij}]$ be an m by n matrix, and let $\mathbf{B} = [b_{ij}]$ be a k by l matrix. The Kronecker product is defined

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \dots & a_{2n}\mathbf{B} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{bmatrix}$$

and has dimension mk by nl .

It can be shown that

Theorem 5.4 (Kronecker and *vec* of a product). *Let \mathbf{A} , \mathbf{B} and \mathbf{C} be conformable matrices as needed. Then*

$$\text{vec}(\mathbf{ABC}) = (\mathbf{C}' \otimes \mathbf{A}) \text{vec } \mathbf{B}$$

$$\Gamma_s \neq \Gamma_{-s}$$

but⁴

$$\Gamma_s = \Gamma'_{-s}.$$

In contrast, the autocovariances of stationary scalar processes satisfy $\gamma_s = \gamma_{-s}$. Computing the autocovariances uses the backward substitution form so that

$$\begin{aligned}\Gamma_s &= E[(\mathbf{y}_t - \mu)(\mathbf{y}_{t-s} - \mu)'] = E\left[\left(\sum_{i=0}^{\infty} \Phi_1^i \boldsymbol{\varepsilon}_{t-i}\right)\left(\sum_{i=0}^{\infty} \Phi_1^i \boldsymbol{\varepsilon}_{t-s-i}\right)'\right] \\ &= E\left[\left(\sum_{i=0}^{s-1} \Phi_1^i \boldsymbol{\varepsilon}_{t-i}\right)\left(\sum_{i=0}^{\infty} \Phi_1^i \boldsymbol{\varepsilon}_{t-s-i}\right)'\right]\end{aligned}\quad (5.12)$$

$$\begin{aligned}&\quad + E\left[\left(\sum_{i=0}^{\infty} \Phi_1^s \Phi_1^i \boldsymbol{\varepsilon}_{t-s-i}\right)\left(\sum_{i=0}^{\infty} \Phi_1^i \boldsymbol{\varepsilon}_{t-s-i}\right)'\right] \\ &= \mathbf{0} + \Phi_1^s E\left[\left(\sum_{i=0}^{\infty} \Phi_1^i \boldsymbol{\varepsilon}_{t-s-i}\right)\left(\sum_{i=0}^{\infty} \Phi_1^i \boldsymbol{\varepsilon}_{t-s-i}\right)'\right] \\ &= \Phi_1^s V[\mathbf{y}_t]\end{aligned}\quad (5.13)$$

and

$$\begin{aligned}\Gamma_{-s} &= E[(\mathbf{y}_t - \mu)(\mathbf{y}_{t+s} - \mu)'] = E\left[\left(\sum_{i=0}^{\infty} \Phi_1^i \boldsymbol{\varepsilon}_{t-i}\right)\left(\sum_{i=0}^{\infty} \Phi_1^i \boldsymbol{\varepsilon}_{t+s-i}\right)'\right] \\ &= E\left[\left(\sum_{i=0}^{\infty} \Phi_1^i \boldsymbol{\varepsilon}_{t-i}\right)\left(\sum_{i=0}^{\infty} \Phi_1^s \Phi_1^i \boldsymbol{\varepsilon}_{t-i}\right)'\right]\end{aligned}\quad (5.14)$$

$$\begin{aligned}&\quad + E\left[\left(\sum_{i=0}^{\infty} \Phi_1^i \boldsymbol{\varepsilon}_{t-i}\right)\left(\sum_{i=0}^{s-1} \Phi_1^i \boldsymbol{\varepsilon}_{t+s-i}\right)'\right] \\ &= E\left[\left(\sum_{i=0}^{\infty} \Phi_1^i \boldsymbol{\varepsilon}_{t-i}\right)\left(\sum_{i=0}^{\infty} \boldsymbol{\varepsilon}'_{t-i} (\Phi'_1)^i (\Phi'_1)^s\right)\right] + \mathbf{0} \\ &= E\left[\left(\sum_{i=0}^{\infty} \Phi_1^i \boldsymbol{\varepsilon}_{t-i}\right)\left(\sum_{i=0}^{\infty} \boldsymbol{\varepsilon}'_{t-i} (\Phi'_1)^i\right)\right] (\Phi'_1)^s \\ &= V[\mathbf{y}_t] (\Phi'_1)^s\end{aligned}\quad (5.15)$$

where $V[\mathbf{y}_t]$ is the symmetric covariance matrix of the VAR. Like most properties of a VAR, the autocovariance function of a VAR(1) closely resembles that of an AR(1): $\gamma_s = \phi_1^{|s|} \sigma^2 / (1 - \phi_1^2) = \phi_1^{|s|} V[y_t]$.

⁴This follows directly from the property of a transpose that if \mathbf{A} and \mathbf{B} are compatible matrices, $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$.

5.2 Companion Form

Any stationary VAR(P) can be rewritten as a VAR(1). Suppose $\{\mathbf{y}_t\}$ follows a VAR(P) process,

$$\mathbf{y}_t = \Phi_0 + \Phi_1 \mathbf{y}_{t-1} + \Phi_2 \mathbf{y}_{t-2} + \dots + \Phi_P \mathbf{y}_{t-P} + \varepsilon_t.$$

By subtracting the mean and stacking P lags of \mathbf{y}_t into a large column vector denoted \mathbf{z}_t , a VAR(P) is equivalently expressed as a VAR(1) using the companion form.

Definition 5.8 (Companion Form of a VAR(P)). Let \mathbf{y}_t follow a VAR(P) given by

$$\mathbf{y}_t = \Phi_0 + \Phi_1 \mathbf{y}_{t-1} + \Phi_2 \mathbf{y}_{t-2} + \dots + \Phi_P \mathbf{y}_{t-P} + \varepsilon_t$$

where ε_t is a vector white noise process and $\mu = \left(\mathbf{I} - \sum_{p=1}^P \Phi_p\right)^{-1} \Phi_0 = E[\mathbf{y}_t]$ is finite. The companion form is

$$\mathbf{z}_t = \Upsilon \mathbf{z}_{t-1} + \xi_t \quad (5.16)$$

where

$$\mathbf{z}_t = \begin{bmatrix} \mathbf{y}_t - \mu \\ \mathbf{y}_{t-1} - \mu \\ \vdots \\ \mathbf{y}_{t-P+1} - \mu \end{bmatrix}, \quad (5.17)$$

$$\Upsilon = \begin{bmatrix} \Phi_1 & \Phi_2 & \Phi_3 & \dots & \Phi_{P-1} & \Phi_P \\ \mathbf{I}_k & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_k & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{I}_k & \mathbf{0} \end{bmatrix} \quad (5.18)$$

and

$$\xi_t = \begin{bmatrix} \varepsilon_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}, E[\xi_t \xi_t'] = \begin{bmatrix} \Sigma & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \end{bmatrix}.. \quad (5.19)$$

The properties of a VAR(P) are identical to that of its companion form VAR(1). For example, VAR(P) is covariance stationary if all of the eigenvalues of Υ - there are $k \times P$ of them - are less than one in absolute value (modulus if complex).⁵

5.3 Empirical Examples

Two examples from the macrofinance literature are used throughout this chapter to illustrate the application of VARs.

⁵The companion form is also useful when working with univariate AR(P) models. An AR(P) can be equivalently expressed as a VAR(1), which simplifies computing properties such as the long-run variance and autocovariances.

5.3.1 Example: The interaction of stock and bond returns

Stocks and bonds are often thought to hedge one another. VARs provide a simple method to determine whether their returns are linked through time. Consider the VAR(1)

$$\begin{bmatrix} VWM_t \\ TERM_t \end{bmatrix} = \begin{bmatrix} \phi_{01} \\ \phi_{02} \end{bmatrix} + \begin{bmatrix} \phi_{11,1} & \phi_{12,1} \\ \phi_{21,1} & \phi_{22,1} \end{bmatrix} \begin{bmatrix} VWM_{t-1} \\ TERM_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}$$

where VWM_t is the return on the value-weighted-market portfolio and $TERM_t$ is the return on a portfolio that is long the 10-year and short the 1-year U.S. government bond. The VAR contains a model for stock returns

$$VWM_t = \phi_{01} + \phi_{11,1}VWM_{t-1} + \phi_{12,1}TERM_{t-1} + \varepsilon_{1,t}$$

and a model for the return on the term premium,

$$TERM_t = \phi_{01} + \phi_{21,1}VWM_{t-1} + \phi_{22,1}TERM_{t-1} + \varepsilon_{2,t}.$$

Since these models do not share any parameters, the coefficient can be estimated equation-by-equation using OLS.⁶ A VAR(1) is estimated using monthly return data (multiplied by 12) for the VWM from CRSP and the 10-year constant maturity treasury yield from FRED covering the period February 1962 until December 2018.⁷

$$\begin{bmatrix} VWM_t \\ TERM_t \end{bmatrix} = \begin{bmatrix} 0.801 \\ (0.000) \\ 0.232 \\ (0.041) \end{bmatrix} + \begin{bmatrix} 0.059 & 0.166 \\ (0.122) & (0.004) \\ -0.104 & 0.116 \\ (0.000) & (0.002) \end{bmatrix} \begin{bmatrix} VWM_{t-1} \\ TERM_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}$$

The p-value of each coefficient is reported in parenthesis. The estimates indicate that stock returns are not predictable using past stock returns but are predictable using the returns on the lagged term premium: positive returns on the term premium lead increase expected returns in stocks. In contrast, positive returns in equities decrease the expected return on the term premium. The annualized long-run mean can be computed from the estimated parameters as

$$12 \times \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 0.059 & 0.166 \\ -0.104 & 0.116 \end{bmatrix} \right)^{-1} \begin{bmatrix} 0.801 \\ 0.232 \end{bmatrix} = \begin{bmatrix} 10.558 \\ 1.907 \end{bmatrix}.$$

These model-based estimates are similar to the sample averages of returns of 10.57 and 1.89 for VWM and $TERM$, respectively.

5.3.2 Example: Monetary Policy VAR

VARs are widely used in macrofinance to model closely related macroeconomic variables. This example uses a 3-variable VAR containing the unemployment rate, the effective federal funds rate, which is the rate that banks use to lend to each other, and inflation. Inflation is measured using the implicit

⁶Theoretical motivations often lead to cross-parameter equality restrictions in VARs. These models cannot be estimated equation-by-equation. A VAR subject to linear equality restrictions can be estimated using a system OLS estimator.

⁷The yields of the bonds are converted to prices, and then returns are computed as the log difference of the prices plus accrued interest.

Raw Data			
	$\Delta \ln \text{UNEMP}_{t-1}$	FF_{t-1}	ΔINF_{t-1}
$\Delta \ln \text{UNEMP}_t$	0.624 (0.000)	0.015 (0.001)	0.016 (0.267)
FF_t	-0.816 (0.000)	0.979 (0.000)	-0.045 (0.317)
ΔINF_t	-0.501 (0.010)	-0.009 (0.626)	-0.401 (0.000)

Standardized Series			
	$\Delta \ln \text{UNEMP}_{t-1}$	FF_{t-1}	ΔINF_{t-1}
$\Delta \ln \text{UNEMP}_t$	0.624 (0.000)	0.153 (0.001)	0.053 (0.267)
FF_t	-0.080 (0.000)	0.979 (0.000)	-0.015 (0.317)
ΔINF_t	-0.151 (0.010)	-0.028 (0.626)	-0.401 (0.000)

Table 5.1: Parameter estimates from the monetary policy VAR. The top panel contains estimates using original, unmodified values while the bottom panel contains estimates from data standardized to have unit variance. While the magnitudes of many coefficients change, the p-values and the eigenvalues of the parameter matrices are identical, and the parameters are roughly comparable since the series have the same variance.

GDP price deflator. Two of the three variables, the unemployment and inflation rates, appear to be nonstationary when tested using an ADF test, and so are differenced.⁸

Using a VAR(1) specification, the model can be described

$$\begin{bmatrix} \Delta \text{UNEMP}_t \\ \text{FF}_t \\ \Delta \text{INF}_t \end{bmatrix} = \Phi_0 + \Phi_1 \begin{bmatrix} \Delta \text{UNEMP}_{t-1} \\ \text{FF}_{t-1} \\ \Delta \text{INF}_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \end{bmatrix}.$$

Two sets of parameters are presented in Table 5.1. The top panel contains estimates using non-scaled data. The bottom panel contains estimates from data where each series is standardized to have unit variance. Standardization produces coefficients that have comparable magnitudes. Despite this transformation and very different parameter estimates, the p-values remain unchanged since OLS t -stats are invariant to rescalings of this type. The eigenvalues of the two parameter matrices are identical, and so the estimate of the persistence of the process is not affected by standardizing the data.

⁸All three series, UNRATE (unemployment), DFF (Federal Funds), and GDPDEF (deflator), are available in FRED. The unemployment and Federal Funds rates are aggregated to quarterly by taking the mean of all observations within a quarter. The inflation rate is computed from the deflator as $400\ln(\text{GDPDEF}_t/\text{GDPDEF}_{t-1})$.

5.4 VAR forecasting

Constructing forecasts of a vector time series is identical to constructing the forecast from a single time series. h -step forecasts are recursively constructed starting with $E_t[\mathbf{y}_{t+1}]$, using $E_t[\mathbf{y}_{t+1}]$ to construct $E_t[\mathbf{y}_{t+2}]$, and continuing until $E_t[\mathbf{y}_{t+h}]$.

Recall that the h -step ahead forecast, $\hat{y}_{t+h|t}$ in an AR(1) is

$$E_t[y_{t+h}] = \sum_{j=0}^{h-1} \phi_1^j \phi_0 + \phi_1^h y_t.$$

The h -step ahead forecast of a VAR(1), $\hat{\mathbf{y}}_{t+h|t}$, has the same structure, and is

$$E_t[\mathbf{y}_{t+h}] = \sum_{j=0}^{h-1} \Phi_1^j \Phi_0 + \Phi_1^h \mathbf{y}_t.$$

This formula can be used to produce multi-step forecast of any VAR using the companion form.

In practice, it is simpler to compute the forecasts using the deviations form of the VAR since it includes no intercept,

$$\tilde{\mathbf{y}}_t = \Phi_1 \tilde{\mathbf{y}}_{t-1} + \Phi_2 \tilde{\mathbf{y}}_{t-2} + \dots + \Phi_P \tilde{\mathbf{y}}_{t-P} + \boldsymbol{\varepsilon}_t,$$

where $\mu = (\mathbf{I}_k - \Phi_1 - \dots - \Phi_P)^{-1} \Phi_0$ and $\tilde{\mathbf{y}}_t = \mathbf{y}_t - \mu$ are mean 0. The h -step forecasts from the deviations form are computed using the recurrence

$$E_t[\tilde{\mathbf{y}}_{t+h}] = \Phi_1 E_t[\tilde{\mathbf{y}}_{t+h-1}] + \Phi_2 E_t[\tilde{\mathbf{y}}_{t+h-2}] + \dots + \Phi_P E_t[\tilde{\mathbf{y}}_{t+h-P}].$$

starting at $E_t[\tilde{\mathbf{y}}_{t+1}]$. Using the forecast of $E_t[\tilde{\mathbf{y}}_{t+h}]$, the h -step ahead forecast of \mathbf{y}_{t+h} is constructed by adding the long-run mean, $E_t[\mathbf{y}_{t+h}] = \mu + E_t[\tilde{\mathbf{y}}_{t+h}]$.

5.4.1 Example: Monetary Policy VAR

Forecasts from VARs incorporate information beyond the history of a single time series. Table 5.2 contains the relative Mean Square Error of out-of-sample forecasts for the three variables in the policy VAR. Each set of forecasts is produced by recursively estimating model parameters using a minimum of 50% of the available sample. Forecasts are produced for up to 8 quarters ahead. Each series is also forecast using a univariate AR model.

The out-of-sample MSE of the forecasts from a model is defined

$$\text{MSE} = \frac{1}{T-h-R} \sum_{t=R}^{T-h} (y_{t+h} - \hat{y}_{t+h|t})^2$$

where R is the size of the initial in-sample period, y_{t+h} is the realization of the variable in period $t+h$, and $\hat{y}_{t+h|t}$ is the h -step ahead forecast produced at time t . The relative MSE is defined as

$$\text{Relative MSE} = \frac{\text{MSE}}{\text{MSE}_{bm}}$$

Horizon	Series	VAR		AR	
		Restricted	Unrestricted	Restricted	Unrestricted
1	Unemployment	0.522	0.520	0.507	0.507
	Fed. Funds Rate	0.887	0.903	0.923	0.933
	Inflation	0.869	0.868	0.839	0.840
2	Unemployment	0.716	0.710	0.717	0.718
	Fed. Funds Rate	0.923	0.943	<i>1.112</i>	<i>1.130</i>
	Inflation	<i>1.082</i>	<i>1.081</i>	<i>1.031</i>	<i>1.030</i>
4	Unemployment	0.872	0.861	0.937	0.940
	Fed. Funds Rate	0.952	0.976	<i>1.082</i>	<i>1.109</i>
	Inflation	<i>1.000</i>	0.999	0.998	0.998
8	Unemployment	0.820	0.806	0.973	0.979
	Fed. Funds Rate	0.974	<i>1.007</i>	<i>1.062</i>	<i>1.110</i>
	Inflation	<i>1.001</i>	1.000	0.998	0.997

Table 5.2: Relative out-of-sample Mean Square Error for forecasts between 1 and 8-quarters ahead. The benchmark model is a constant for the unemployment rate and the inflation rate and a random walk for the Federal Funds rate. Model parameters are recursively estimated, and forecasts are produced once 50% of the available sample. Model order is selected using the BIC.

where MSE_{bm} is the out-of-sample MSE of a benchmark model. The Federal Funds rate is modeled in levels, and so the benchmark model is a random walk. The other two series are differenced, and so use the historical mean (an AR(0)) as the benchmark model. The number of lags in either the VAR or the AR is selected by minimizing the BIC (see Section 5.5).

Each model is estimated using two methods, the standard estimator and a restricted estimator where the long-run mean forced to match the in-sample mean. The restricted model is estimated in two steps. First, the sample mean is subtracted, and then the model is estimated without a constant. The forecasts are then constructed using the sample mean plus the forecast of the demeaned data. The two-step estimator ensures that the model mean reverts to the historical average. The unrestricted model jointly estimates the intercept with the parameters that capture the dynamics and so does not revert (exactly) to the sample mean even over long horizons. These two method can produce qualitatively different forecasts in persistent time series due to differences in the average values of the data used as lags ($\bar{\mathbf{y}}_{t-j} = (T-P)^{-1} \sum_{t=P-j+1}^{T-j} \mathbf{y}_t$ for $j = 1, 2, \dots, P$) and the average value of the contemporaneous values ($\bar{\mathbf{y}}_t = (T-P)^{-1} \sum_{t=P+1}^T \mathbf{y}_t$). The two-step estimator uses the same mean value for both, $\bar{\mathbf{y}} = T^{-1} \sum_{t=1}^P \mathbf{y}_t$.

The VAR performs well in this forecasting problem. It produced the lowest MSE in 7 of 12 horizon-series combinations. When it is not the best model, it performs only slightly worse than the autoregression. Ultimately, the choice of a model to use in forecasting applications – either multivariate or univariate – is an empirical question that is best answered using in-sample analysis and pseudo-out-of-sample forecasting.

5.5 Estimation and Identification

Understanding the dependence structure in VAR models requires additional measures of cross-variable relationships. The cross-correlation function (CCF) and partial cross-correlation function (PCCF) extend the autocorrelation and partial autocorrelation functions used to identify the model order in a single time series.

Definition 5.9 (Cross-correlation). The s^{th} cross correlations between two covariance stationary series $\{x_t\}$ and $\{y_t\}$ are defined

$$\rho_{xy,s} = \frac{E[(x_t - \mu_x)(y_{t-s} - \mu_y)]}{\sqrt{V[x_t]V[y_t]}} \quad (5.20)$$

and

$$\rho_{yx,s} = \frac{E[(y_t - \mu_y)(x_{t-s} - \mu_x)]}{\sqrt{V[x_t]V[y_t]}} \quad (5.21)$$

where the order of the indices indicates the lagged variable, $E[y_t] = \mu_y$ and $E[x_t] = \mu_x$.

Cross-correlations, unlike autocorelations, are not symmetric in the order of the arguments. Partial cross-correlations are defined using a similar extension of partial autocorrelation as the correlation between x_t and y_{t-s} controlling for $y_{t-1}, \dots, y_{t-(s-1)}$.

Definition 5.10 (Partial Cross-correlation). The partial cross-correlations between two covariance stationary series $\{x_t\}$ and $\{y_t\}$ are defined as the population values of the coefficients $\varphi_{xy,s}$ in the regression

$$x_t = \phi_0 + \phi_{x1}x_{t-1} + \dots + \phi_{xs-1}x_{t-(s-1)} + \phi_{y1}y_{t-1} + \dots + \phi_{ys-1}y_{t-(s-1)} + \varphi_{xy,s}y_{t-s} + \varepsilon_{x,t} \quad (5.22)$$

and $\varphi_{yx,s}$ in the regression

$$y_t = \phi_0 + \phi_{y1}y_{t-1} + \dots + \phi_{ys-1}y_{t-(s-1)} + \phi_{x1}x_{t-1} + \dots + \phi_{xs-1}x_{t-(s-1)} + \varphi_{yx,s}x_{t-s} + \varepsilon_{y,t} \quad (5.23)$$

where the order of the indices indicates which lagged variable. In a k -variable VAR, the PCCF of series i with respect to series j is the population value of the coefficient $\varphi_{y_i y_j s}$ in the regression

$$y_{it} = \phi_0 + \phi'_1 y_{t-1} + \dots + \phi'_{s-1} + \varphi_{y_i y_j s} y_{jt-s} + \varepsilon_i$$

where ϕ_j are 1 by k vectors of parameters.

The controls in the s^{th} partial cross-correlation are included variables in a VAR($s-1$). If the data are generated by a VAR(P), then the s^{th} partial cross-correlation is 0 whenever $s > P$. This behavior is analogous to the behavior of the PACF in an AR(P) model. The PCCF is a useful diagnostic to identify the order of a VAR and for verifying the order of estimated models when applied to residuals.

Figure 5.1 plots 1,000 simulated data points from a high-order bivariate VAR. One component of the VAR follows a HAR(22) process with no spillovers from the other component. The second component is substantially driven by both spillovers from the HAR and its own innovation. The complete specification of the VAR(22) is

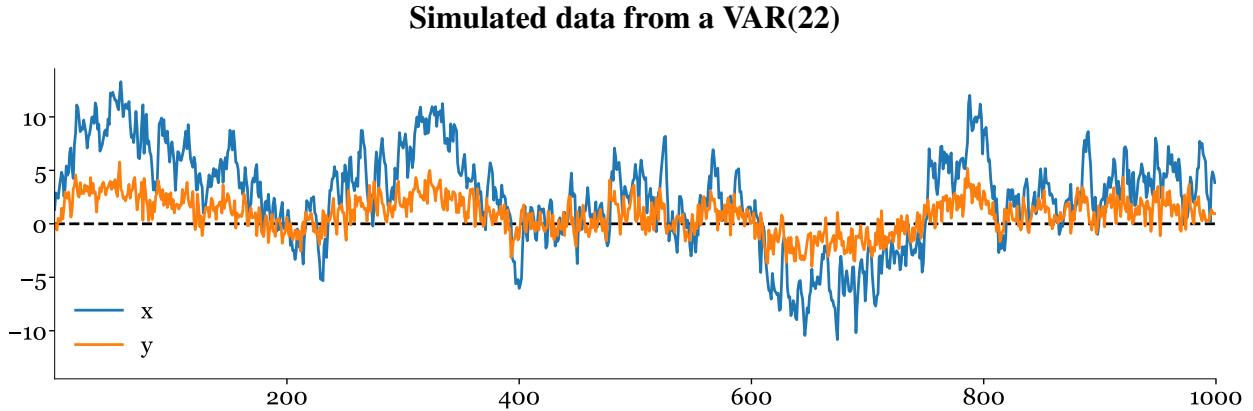


Figure 5.1: Simulated data from the VAR(22) in eq. (5.24). Both processes are stationary but highly persistent and have a high degree of comovement.

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} 0.5 & 0.9 \\ 0 & 0.47 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \sum_{i=2}^5 \begin{bmatrix} 0 & 0 \\ 0 & 0.06 \end{bmatrix} \begin{bmatrix} x_{t-i} \\ y_{t-i} \end{bmatrix} + \sum_{j=6}^{22} \begin{bmatrix} 0 & 0 \\ 0 & 0.01 \end{bmatrix} \begin{bmatrix} x_{t-j} \\ y_{t-j} \end{bmatrix} + \begin{bmatrix} \varepsilon_{x,t} \\ \varepsilon_{y,t} \end{bmatrix}. \quad (5.24)$$

Figure 5.2 contains plots of the theoretical ACF and CCF (cross-correlation function) of this VAR. Both ACFs and CCFs indicate that the series are highly persistent. They also show that both variables are a strong predictor of either at any lag since the squared correlation can be directly interpretable as an R^2 . Figure 5.3 contains plots of the partial auto- and cross-correlation function. These are markedly different from the ACFs and CCFs. The PACF and PCCF of x both cut off after one lag. This happens since x has 0 coefficients on all lagged values after the first. The PACF and PCCF of y are more complex. The PACF resembles the step-function of the coefficients in the HAR model. It cuts off sharply after 22 lags since this is the order of the VAR. The PCCF of y is also non-zero for many lags, and only cuts off after 21. The reduction in the cut-off is due to the structure of the VAR where x is only exposed to the lagged value of y at the first lag, and so the dependence is reduced by one.

These new definitions enable the key ideas of the Box-Jenkins methodology to be extended to vector processes. While this extension is technically possible, using the ACF, PACF, CCF, and PCCF to determine the model lag length is difficult. The challenge of graphical identification of the order is especially daunting in specifications with more than two variables since there are many dependence measures to inspect – a k -dimensional stochastic process has $2(k^2 - k)$ distinct auto- and cross-correlation functions.

The standard approach is to adopt the approach advocated in Sims (1980). The VAR specification should include all variables that theory indicates are relevant, and the lag length should be chosen so that the model has a high likelihood of capturing all of the dynamics. Once the maximum value of the lag length is chosen, a general-to-specific search can be conducted to reduce the model order, or an information criterion can be used to select an appropriate lag length. In a VAR, the Akaike IC, Hannan and Quinn (1979) IC and the Schwarz/Bayesian IC are

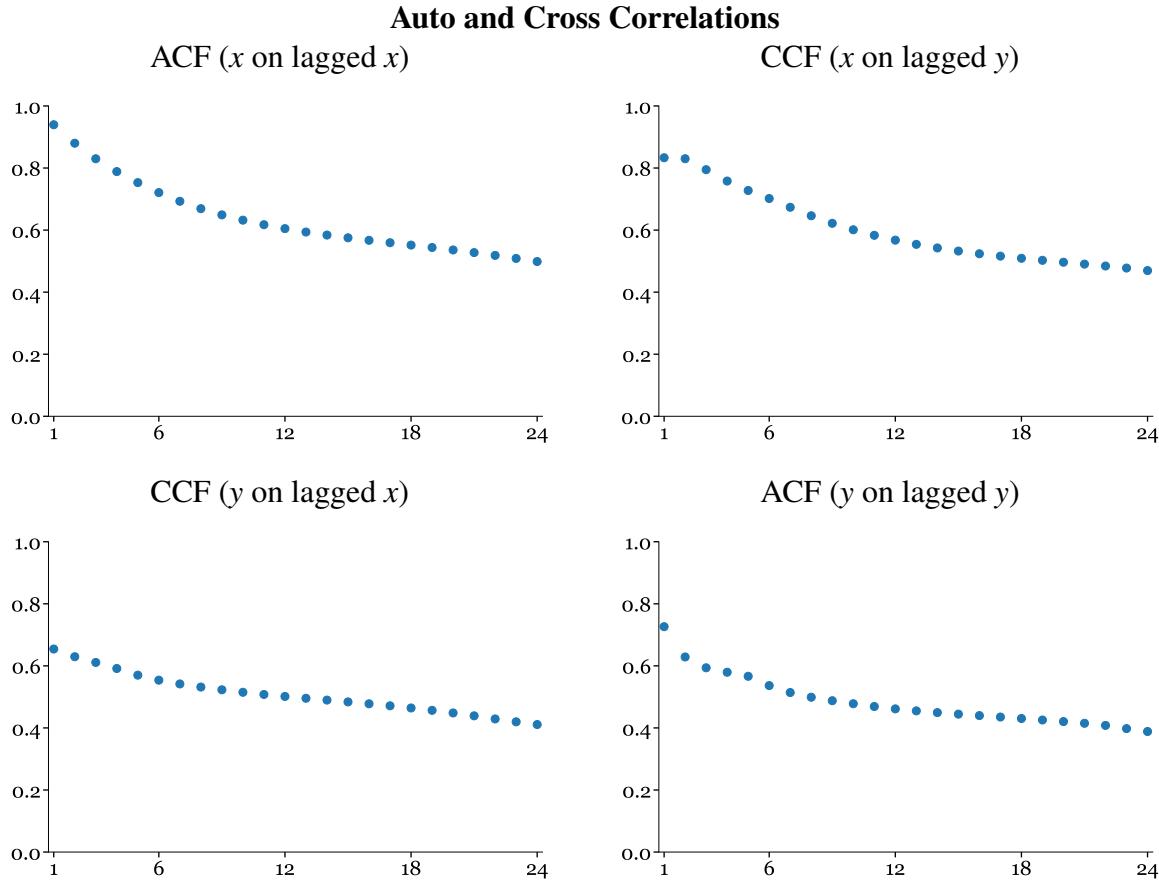


Figure 5.2: The four panels contain the ACFs and CCFs of the VAR(22) process in eq. (5.24).

$$\begin{aligned} \text{AIC: } & \ln |\hat{\Sigma}(P)| + k^2 P \frac{2}{T} \\ \text{HQIC: } & \ln |\hat{\Sigma}(P)| + k^2 P \frac{2 \ln \ln T}{T} \\ \text{BIC: } & \ln |\hat{\Sigma}(P)| + k^2 P \frac{\ln T}{T} \end{aligned}$$

where $\hat{\Sigma}(P)$ is the covariance of the residuals estimated using a $\text{VAR}(P)$ and $|\cdot|$ is the determinant.⁹ All models must use the same values on the left-hand-side irrespective of the lags included when choosing the lag length. In practice, it is necessary to adjust the sample when estimating the parameters of models with fewer lags than the maximum allowed. For example, when comparing models with up to 2 lags, the largest model is estimated by fitting observations $3, 4, \dots, T$ since two lags are lost when constructing the right-hand-side variables. The 1-lag model should also fit observations $3, 4, \dots, T$ and so observation 1 is excluded from the model since it is not needed as a lagged variable.

⁹ $\ln |\hat{\Sigma}|$ is, up to an additive constant, the Gaussian log-likelihood divided by T . These three information criteria are all special cases of the usual information criteria for log-likelihood models which take the form $-L + P_{IC}$ where P_{IC} is the penalty which depends on the number of estimated parameters in the model and the information criterion.

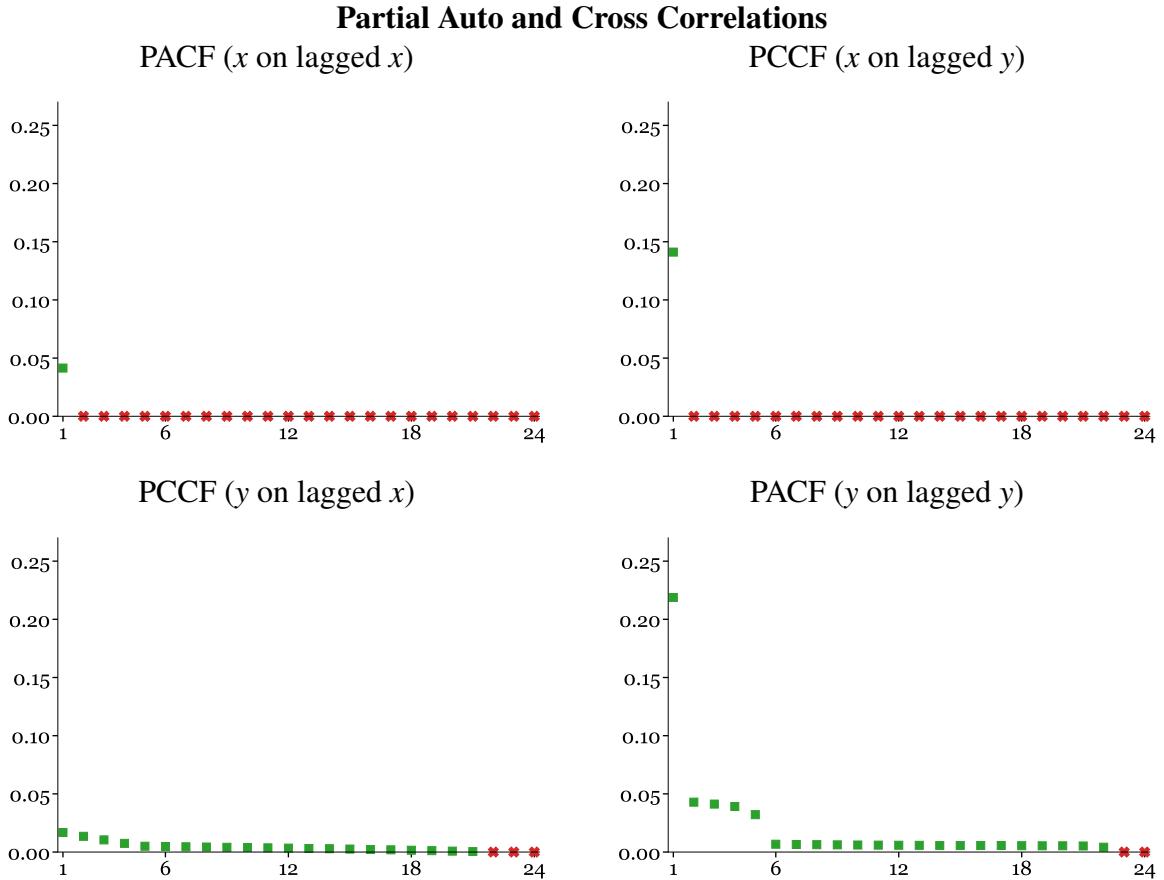


Figure 5.3: The four panels contain the PACFs and PCCFs of the VAR(22) process in eq. (5.24). Values marked with a red **x** are exactly 0.

The lag length should be chosen to minimize one of the criteria. The BIC has the most substantial penalty term and so always chooses a (weakly) smaller model than the HQIC. The AIC has the smallest penalty, and so selects the largest model of the three ICs. Ivanov and Kilian (2005) recommend the AIC for monthly models and the HQIC for quarterly models unless the sample size is less than 120 quarters. In short samples, the BIC is preferred. Their recommendation is based on the accuracy of the impulse response function, and so may not be ideal in other applications, e.g., forecasting.

Alternatively, a likelihood ratio test can be used to test whether two specifications are equivalent. The LR test statistic is

$$(T - P_2 k^2) (\ln |\hat{\Sigma}(P_1)| - \ln |\hat{\Sigma}(P_2)|) \stackrel{A}{\sim} \chi_{(P_2 - P_1)k^2}^2,$$

where P_1 is the number of lags in the restricted (smaller) model, P_2 is the number of lags in the unrestricted (larger) model and k is the dimension of \mathbf{y}_t . Since model 1 is a restricted version of model 2, its covariance is larger and so this statistic is always positive. The $-P_2 k^2$ term in the log-likelihood is a degree of freedom correction that generally improves small-sample performance of the test. Ivanov and Kilian (2005) recommend against using sequential likelihood ratio testing in lag length selection.

Lag Length	AIC	HQIC	BIC	LR	P-val
0	4.014	3.762	3.605	925	0.000
1	0.279	0.079	0.000▼▲	39.6	0.000
2	0.190	0.042	0.041	40.9	0.000
3	0.096	0.000▼	0.076	29.0	0.001
4	0.050▼	0.007	0.160	7.34	0.602▼
5	0.094	0.103	0.333	29.5	0.001
6	0.047	0.108	0.415	13.2	0.155
7	0.067	0.180	0.564	32.4	0.000
8	0.007	0.172▲	0.634	19.8	0.019
9	0.000▲	0.217	0.756	7.68	0.566▲
10	0.042	0.312	0.928	13.5	0.141
11	0.061	0.382	1.076	13.5	0.141
12	0.079	0.453	1.224	—	—

Table 5.3: Normalized values for the AIC, HQIC, and BIC in a Monetary Policy VAR. The information criteria are normalized by subtracting the smallest value from each column. The LR and P-value in each row are for a test with the null that the coefficient on lag $l + 1$ are all zero ($H_0 : \Phi_{l+1} = \mathbf{0}$) and the alternative $H_1 : \Phi_{l+1} \neq \mathbf{0}$. Values marked with ▼ indicate the lag length selected using a specific-to-general search. Values marked with ▲ indicate the lag length selected using general-to-specific.

5.5.1 Example: Monetary Policy VAR

The Monetary Policy VAR is used to illustrate lag length selection. The information criteria, log-likelihoods, and p-values from the LR tests are presented in Table 5.3. This table contains the AIC, HQIC, and BIC values for lags 0 (no dynamics) through 12 as well as likelihood ratio test results for testing l lags against $l + 1$. Note that the LR and P-value corresponding to lag l test the null that the fit using l lags is equivalent to the fit using $l + 1$ lags. Using the AIC, 9 lags produces the smallest value and is selected in a general-to-specific search. A specific-to-general search stops at 4 lags since the AIC of 5 lags is larger than the AIC of 4. The HQIC chooses 3 lags in a specific-to-general search and 9 in a general-to-specific search. The BIC selects a single lag irrespective of the search direction. A general-to-specific search using the likelihood ratio chooses 9 lags, and a hypothesis-test-based specific-to-general procedure chooses 4. The specific-to-general stops at 4 lags since the null $H_0 : P = 4$ tested against the alternative that $H_1 : P = 5$ has a p-value of 0.602, which indicates that these models provide similar fits of the data.

Finally, the information criteria are applied in a “global search” that evaluates models using every combination of lags up to 12. This procedure fits a total of 4096 VARs (which only requires a few seconds on a modern computer), and the AIC, HQIC, and the BIC are computed for each.¹⁰ Using this methodology, the AIC search selected lags 1–3 and 7–9, the HQIC selects lags 1–3, 6, and 8, and the BIC continues to select a parsimonious model that includes only the first lag. Search procedures of this type are computationally viable for checking up to 20 lags.

¹⁰For a maximum lag length of L , 2^L models must be estimated.

5.6 Granger causality

Granger causality (GC, also known as *prima facia* causality) is the first concept exclusive to vector analysis. GC is the standard method to determine whether one variable is useful in predicting another and evidence of Granger causality it is a good indicator that a VAR, rather than a univariate model, is needed.

Granger causality is defined in the negative.

Definition 5.11 (Granger causality). A scalar random variable x_t does not Granger cause y_t if $E[y_t|x_{t-1}, y_{t-1}, x_{t-2}, y_{t-2}, \dots] = E[y_t|y_{t-1}, y_{t-2}, \dots]$.¹¹ That is, x_t does not Granger cause y_t if the forecast of y_t is the same whether conditioned on past values of x_t or not.

Granger causality can be simply illustrated in a bivariate VAR.

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} \phi_{11,1} & \phi_{12,1} \\ \phi_{21,1} & \phi_{22,1} \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \phi_{11,2} & \phi_{12,2} \\ \phi_{21,2} & \phi_{22,2} \end{bmatrix} \begin{bmatrix} x_{t-2} \\ y_{t-2} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}$$

In this model, if $\phi_{21,1} = \phi_{21,2} = 0$ then x_t does not Granger cause y_t . Note that x_t not Granger causing y_t says nothing about whether y_t Granger causes x_t .

An important limitation of GC is that it does not account for indirect effects. For example, suppose x_t and y_t are both Granger caused by z_t . x_t is likely to Granger cause y_t in a model that omits z_t if $E[y_t|y_{t-1}, x_{t-1}, \dots] \neq E[y_t|y_{t-1}, z_{t-1}, \dots]$ even though $E[y_t|y_{t-1}, z_{t-1}, x_{t-1}, \dots] = E[y_t|y_{t-1}, z_{t-1}, \dots]$.

Testing

Testing Granger causality in a VAR(P) is implemented using a likelihood ratio test. In the VAR(P),

$$\mathbf{y}_t = \Phi_0 + \Phi_1 \mathbf{y}_{t-1} + \Phi_2 \mathbf{y}_{t-2} + \dots + \Phi_P \mathbf{y}_{t-P} + \boldsymbol{\varepsilon}_t,$$

$y_{j,t}$ does not Granger cause $y_{i,t}$ if $\phi_{ij,1} = \phi_{ij,2} = \dots = \phi_{ij,P} = 0$. The likelihood ratio test statistic for testing the null $H_0 : \phi_{ij,m} = 0, \forall m \in \{1, 2, \dots, P\}$ against the alternative $H_1 : \phi_{ij,m} \neq 0 \exists m \in \{1, 2, \dots, P\}$ is

$$(T - (Pk^2 - k)) (\ln |\hat{\Sigma}_r| - \ln |\hat{\Sigma}_u|) \stackrel{A}{\sim} \chi_P^2$$

where Σ_r is the estimated residual covariance when the null of no Granger causation is imposed ($H_0 : \phi_{ij,1} = \phi_{ij,2} = \dots = \phi_{ij,P} = 0$) and Σ_u is the estimated covariance in the unrestricted VAR(P).¹²

5.6.1 Example: Monetary Policy VAR

The monetary policy VAR is used to illustrate testing Granger causality. Table 5.4 contains the results of Granger causality tests in the monetary policy VAR with three lags (as chosen by the HQIC). Tests of a variable causing itself have been omitted since these are not informative about the need for a

¹¹Technically, this definition is for Granger causality in the mean. Other definition exist for Granger causality in the variance (replace conditional expectation with conditional variance) and distribution (replace conditional expectation with conditional distribution).

¹²The multiplier in the test is a degree of freedom adjusted factor. There are T data points, and there are $Pk^2 - k$ parameters in the restricted model.

Exclusion	Fed. Funds Rate		Inflation		Unemployment	
	P-val	Stat	P-val	Stat	P-val	Stat
Fed. Funds Rate	–	–	0.001	13.068	0.014	8.560
Inflation	0.001	14.756	–	–	0.375	1.963
Unemployment	0.000	19.586	0.775	0.509	–	–
All	0.000	33.139	0.000	18.630	0.005	10.472

Table 5.4: Tests of Granger causality. This table contains tests where the variable on the left-hand side is excluded from the regression for the variable along the top. Since the null is no GC, rejection indicates a relationship between past values of the variable on the left and contemporaneous values of variables on the top.

multivariate model. The table contains tests whether the variables in the left-hand column Granger Cause the variables labeled across the top. Each row contains a p-value indicating significance using standard test sizes (5 or 10%), and so each variable causes at least one other variable. Column-by-column examination demonstrated that every variable is caused by at least one other variable. The final row labeled All tests the null that a univariate model performs as well as a multivariate model by restricting all variable other than the target to have zero coefficients. This test further confirms that the VAR is required for each component.

5.7 Impulse Response Functions

In the univariate world, the $\text{MA}(\infty)$ representation of an ARMA is sufficient to understand how a shock decays. When analyzing vector data, this is no longer the case. A shock to one series has an immediate effect on that series, but it can also affect the other variables in the system which, in turn, feed back into the original variable. It is not possible to visualize the propagation of a shock using only the estimated parameters in a VAR. Impulse response functions simplify this task by providing a visual representation of shock propagation.

5.7.1 Defined

Definition 5.12 (Impulse Response Function). The impulse response function of y_i , an element of \mathbf{y} , with respect to a shock in ε_j , an element of ε , for any j and i , is defined as the change in y_{it+s} , $s \geq 0$ for a one standard deviation shock in $\varepsilon_{j,t}$.

This definition is somewhat difficult to parse and the impulse response function easier to understand using the vector moving average (VMA) representation of a VAR.¹³ When \mathbf{y}_t is covariance stationary then it must have a VMA representation,

$$\mathbf{y}_t = \mu + \varepsilon_t + \Xi_1 \varepsilon_{t-1} + \Xi_2 \varepsilon_{t-2} + \dots$$

¹³Recall that a stationary AR(P) can also be transformed into a $\text{MA}(\infty)$. Transforming a stationary VAR(P) into a VMA(∞) is the multivariate time-series analogue.

Using this VMA, the impulse response of y_i with respect to a shock in ε_j at period h is

$$IRF_h = \sigma_j \Xi_h \mathbf{e}_j \quad (5.25)$$

where \mathbf{e}_j is a vector of 0s with 1 in position j , $\mathbf{e}_j = \begin{bmatrix} 0, \dots, 0, 1, 0, \dots, 0 \end{bmatrix}'$ and where σ_j is the standard deviation of ε_j . These impulse responses are then $\{\sigma_j, \sigma_j \Xi_1^{[ii]}, \sigma_j \Xi_2^{[ii]}, \sigma_j \Xi_3^{[ii]}, \dots\}$ if $i = j$ and $\{0, \sigma_j \Xi_1^{[ij]}, \sigma_j \Xi_2^{[ij]}, \sigma_j \Xi_3^{[ij]}, \dots\}$ otherwise where $\Xi_m^{[ij]}$ is the element in row i and column j of Ξ_m . The coefficients of the VMA can be computed from the VAR using the relationship

$$\Xi_j = \Phi_1 \Xi_{j-1} + \Phi_2 \Xi_{j-2} + \dots + \Phi_P \Xi_{j-P}$$

where $\Xi_0 = \mathbf{I}_k$ and $\Xi_m = \mathbf{0}$ for $m < 0$. For example, in a VAR(2),

$$\mathbf{y}_t = \Phi_1 \mathbf{y}_{t-1} + \Phi_2 \mathbf{y}_{t-2} + \varepsilon_t,$$

$\Xi_0 = \mathbf{I}_k$, $\Xi_1 = \Phi_1$, $\Xi_2 = \Phi_1^2 + \Phi_2$, and $\Xi_3 = \Phi_1^3 + \Phi_1 \Phi_2 + \Phi_2 \Phi_1$.

5.7.2 Orthogonal Impulse Response Functions

The previous discussion assumed shocks are uncorrelated so that a shock to component j had no effect on the other components of the error. This assumption is problematic since the shocks are often correlated, and so it is not possible to change one in isolation. The model shocks have covariance $\text{Cov}[\varepsilon_t] = \Sigma$, and so a set of orthogonal shocks can be produced as $\eta_t = \Sigma^{-1/2} \varepsilon_t$. Using these uncorrelated and standardized shocks, the VMA is now

$$\begin{aligned} \mathbf{y}_t &= \mu + \varepsilon_t + \Xi_1 \Sigma^{1/2} \Sigma^{-1/2} \varepsilon_{t-1} + \Xi_2 \Sigma^{1/2} \Sigma^{-1/2} \varepsilon_{t-2} + \dots \\ &= \mu + \Sigma^{1/2} \eta_t + \tilde{\Xi}_1 \eta_{t-1} + \tilde{\Xi}_2 \eta_{t-2} + \dots \end{aligned}$$

where $\tilde{\Xi}_m = \Xi_m \Sigma^{1/2}$. The impulse response for a shock to series j in period h is $\Sigma^{1/2} \mathbf{e}_j$ in period 0,

$$OIRF_h = \tilde{\Xi}_h \mathbf{e}_j \quad (5.26)$$

for $h \geq 1$. If Σ is diagonal, then these impulse responses are identical to the expression in eq. (5.25).

In practice, the Cholesky factor is used as the square root of the covariance matrix. The Cholesky factor is a lower triangular matrix which imposes a de facto ordering to the shocks. For example, if

$$\Sigma = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix},$$

then the Cholesky factor is

$$\Sigma_C^{1/2} = \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix}$$

so that $\Sigma = \Sigma_C^{1/2} (\Sigma_C^{1/2})'$. Shocking element j has an effect of every series the appears after j (j, \dots, k) but not on the first $j-1$ ($1, \dots, j-1$). In some contexts, it is plausible that there is a natural order to the shocks since some series are faster than others.

In the monetary policy VAR, it is commonly assumed that changes in the Federal Funds rate immediately spillover to unemployment and inflation, but that unemployment and inflation only feed-back into the Federal Funds rate with a lag. Similarly, it is commonly assumed that changes in unemployment affect inflation immediately, but that inflation does not have a contemporaneous impact on unemployment. When using the Cholesky factor, the impulse responses depend on the order of the variables in the VAR. Additionally, in many important applications – for example when a VAR includes multiple financial variables – then there is no plausible method to order the shocks since financial variables are likely to react simultaneously to a shock.

The leading alternative to the using the Cholesky factor is to use a Generalized Impulse Response function (Pesaran and Shin, 1998). This method is invariant to the order of the variables since it does not use a matrix square root. The GIRF is justified as the difference measuring between the conditional expectation of \mathbf{y}_{t+h} given shock j is one standard deviation and the conditional expectation of \mathbf{y}_{t+h} ,

$$E_t [\mathbf{y}_{t+h} | \varepsilon_j = \sigma_j] - E_t [\mathbf{y}_{t+h}].$$

When the VAR is driven by normally distributed errors, this expression is

$$GIRF_h = \sigma_j^{-1} \Xi_h \Sigma \mathbf{e}_j. \quad (5.27)$$

The GIRF is equivalently expressed as

$$\Xi_h [\sigma_{1j}, \sigma_{2j}, \dots, \sigma_{kj}]' / \sigma_{jj} \times \sigma_j = \Xi_h [\beta_{1j}, \beta_{2j}, \dots, \beta_{kj}]' \sigma_j$$

where β_{ij} is the population value of the regression coefficient of regressing ε_{it} on ε_{jt} .

5.7.3 Example: Impulse Response in the Monetary Policy VAR

The monetary policy VAR is used to illustrate impulse response functions. Figure 5.4 contains the impulse responses of the three variable to the three shocks. The dotted lines represent two standard deviation confidence intervals. The covariance is factored using the Cholesky, and it is assumed that the shock to the Federal Funds Rate impacts all variables immediately, the shock to the unemployment affects inflation immediately but not the Federal Funds rate, and that the inflation shock has no immediate effect. The unemployment rate is sensitive to changes in the Federal Funds rate, and one standard deviation shock reduces the change (ΔUNEMP_t) in the unemployment rate by up to 0.15% as the impulse evolves.

5.7.4 Confidence Intervals

Impulse response functions, like the parameters of the VAR, are estimated quantities and subject to statistical variation. Confidence bands are used to determine whether an impulse response different from zero. Since the parameters of the VAR are asymptotically normally distributed (as long as it is stationary and the innovations are white noise), the impulse responses also asymptotically normal, which follows as an application of the δ -method. The analytical derivation of the covariance of the impulse response function is tedious (see Section 11.7 in Hamilton (1994) for details). Instead, two computational methods to construct confidence bands of impulse response functions are described: Monte Carlo and bootstrap.

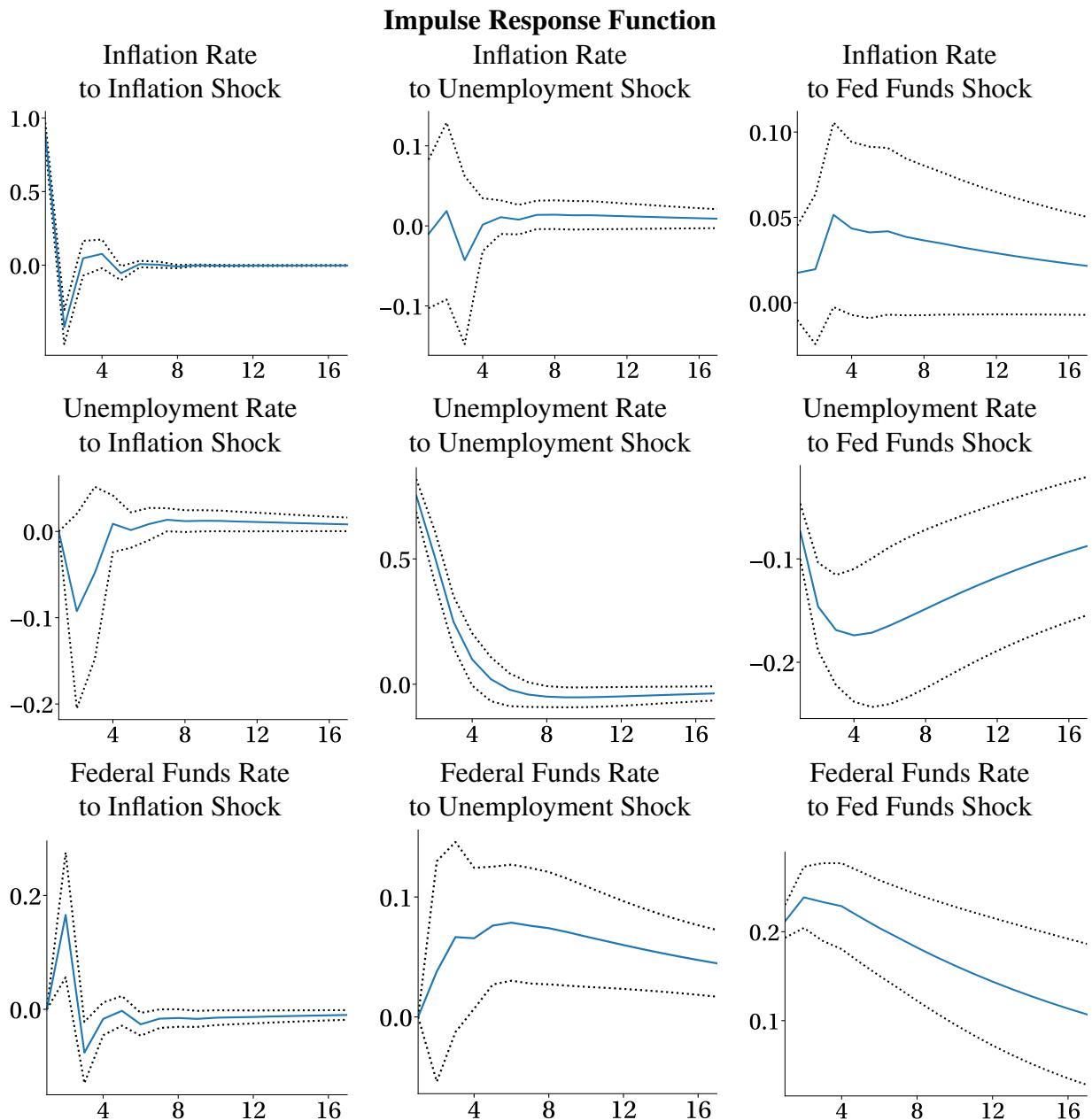


Figure 5.4: Impulse response functions for 16 quarters. The dotted lines represent two standard deviation confidence intervals. The covariance is factored using the Cholesky so that a shock to the Federal Funds rate spills over immediately to the other two variables, an unemployment shock spills over to inflation, and an inflation shock has no immediate effect on the other series.

5.7.4.1 Monte Carlo Confidence Intervals

Monte Carlo confidence intervals come in two forms, one that directly simulates $\hat{\Phi}_i$ from its asymptotic distribution and one that simulates the VAR and draws $\hat{\Phi}_i$ as the result of estimating the unknown parameters in the simulated VAR. The direct sampling method is simple:

1. Compute $\hat{\theta}$ from the data and estimate the covariance matrix $\hat{\Lambda}$ in the asymptotic distribution $\sqrt{T}(\hat{\theta} - \theta) \xrightarrow{A} N(0, \hat{\Lambda})$ where θ is the collection of all model parameters, $\Phi_0, \Phi_1, \dots, \Phi_P$ and Σ .
2. Using $\hat{\theta}$ and $\hat{\Lambda}$, generate simulated values $\hat{\Phi}_{0b}, \hat{\Phi}_{1b}, \dots, \hat{\Phi}_{Pb}$ and $\hat{\Sigma}_b$ from the asymptotic distribution as $\hat{\theta} + \hat{\Lambda}^{1/2} \varepsilon$ where $\varepsilon \stackrel{\text{i.i.d.}}{\sim} N(\mathbf{0}, \mathbf{I}_{k^2(P+1)})$. These are i.i.d. draws from a $N(\hat{\theta}, \hat{\Lambda})$ distribution.
3. Using $\hat{\Phi}_{0b}, \hat{\Phi}_{1b}, \dots, \hat{\Phi}_{Pb}$ and $\hat{\Sigma}_b$, compute the impulse responses $\{\hat{\Xi}_{jb}\}$ where $j = 1, 2, \dots, h$. Save these values.
4. Return to step 2 and compute a total of B impulse responses. Typically B is between 100 and 1000.
5. For each impulse response and each horizon, sort the responses. The 5th and 95th percentile of this distribution are the confidence intervals.

The second Monte Carlo method simulates data assuming the errors are i.i.d. normally distributed, and then uses these values to produce a draw from the joint distribution of the model parameters. This method avoids the estimation of the parameter covariance matrix $\hat{\Lambda}$ in the alternative Monte Carlo method.

1. Compute $\hat{\Phi}$ from the initial data and estimate the residual covariance $\hat{\Sigma}$.
2. Using $\hat{\Phi}$ and $\hat{\Sigma}$, simulate a time-series $\{\tilde{\mathbf{y}}_t\}$ with as many observations as the original data. These can be computed directly using forward recursion

$$\tilde{\mathbf{y}}_t = \hat{\Phi}_0 + \hat{\Phi}_1 \mathbf{y}_{t-1} + \dots + \hat{\Phi}_P \mathbf{y}_{t-P} + \hat{\Sigma}^{1/2} \varepsilon_t$$

where $\varepsilon \stackrel{\text{i.i.d.}}{\sim} N(\mathbf{0}, \mathbf{I}_k)$ are multivariate standard normally distributed. The P initial values are set to a consecutive block of the historical data chosen at random, $\mathbf{y}_\tau, \mathbf{y}_{\tau+1}, \dots, \mathbf{y}_{\tau+P-1}$ for $\tau \in \{1, \dots, T-P\}$.

3. Using $\{\tilde{\mathbf{y}}_t\}$, estimate the model parameters $\hat{\Phi}_{0b}, \hat{\Phi}_{1b}, \dots, \hat{\Phi}_{Pb}$ and $\hat{\Sigma}_b$.
4. Using $\hat{\Phi}_{0b}, \hat{\Phi}_{1b}, \dots, \hat{\Phi}_{Pb}$ and $\hat{\Sigma}_b$, compute the impulse responses $\{\hat{\Xi}_{jb}\}$ where $j = 1, 2, \dots, h$. Save these values.
5. Return to step 2 and compute a total of B impulse responses. Typically B is between 100 and 1000.
6. For each impulse response for each horizon, sort the impulse responses. The 5th and 95th percentile of this distribution are the confidence intervals.

Of these two methods, the former should be preferred since the assumption of i.i.d. normally distributed errors in the latter may be unrealistic, especially when modeling financial data.

5.7.4.2 Bootstrap Confidence Intervals

The bootstrap is a simulation-based method that resamples from the observed data produce a simulated data set. The idea behind this method is simple: if the residuals are realizations of the actual error process, one can use them directly to simulate this distribution rather than making an arbitrary assumption about the error distribution (e.g., i.i.d. normal). The procedure is essentially identical to the second Monte Carlo procedure outlined above:

1. Compute $\hat{\Phi}$ from the initial data and estimate the residuals $\hat{\varepsilon}_t$.
2. Using $\hat{\varepsilon}_t$, compute a new series of residuals $\tilde{\varepsilon}_t$ by sampling, with replacement, from the original residuals. The new series of residuals can be described

$$\{\hat{\varepsilon}_{u_1}, \hat{\varepsilon}_{u_2}, \dots, \hat{\varepsilon}_{u_T}\}$$

where u_i are i.i.d. discrete uniform random variables taking the values $1, 2, \dots, T$. In essence, the new set of residuals is just the old set of residuals reordered with some duplication and omission.¹⁴

3. Using $\hat{\Phi}$ and $\{\hat{\varepsilon}_{u_1}, \hat{\varepsilon}_{u_2}, \dots, \hat{\varepsilon}_{u_T}\}$, simulate a time-series $\{\tilde{y}_t\}$ with as many observations as the original data. These can be computed directly using the VAR

$$\tilde{y}_t = \hat{\Phi}_0 + \hat{\Phi}_1 y_{t-1} + \dots + \hat{\Phi}_P y_{t-P} + \hat{\varepsilon}_{u_t}$$

4. Using $\{\tilde{y}_t\}$, compute estimates of $\hat{\Phi}_{0b}, \hat{\Phi}_{1b}, \dots, \hat{\Phi}_{Pb}$ and $\hat{\Sigma}_b$ from a VAR.
5. Using $\hat{\Phi}_{0b}, \hat{\Phi}_{1b}, \dots, \hat{\Phi}_{Pb}$ and $\hat{\Sigma}_b$, compute the impulse responses $\{\tilde{\Xi}_{jb}\}$ where $j = 1, 2, \dots, h$. Save these values.
6. Return to step 2 and compute a total of B impulse responses. Typically B is between 100 and 1000.
7. For each impulse response for each horizon, sort the impulse responses. The 5th and 95th percentile of this distribution are the confidence intervals.

5.8 Cointegration

Many economic time-series are nonstationarity and so standard VAR analysis which assumes all series are covariance stationary is unsuitable. Cointegration extends stationary VAR models to non-stationary time series. Cointegration analysis also provides a method to characterize the long-run equilibrium of a system of non-stationary variables. Before more formally examining cointegration, consider the consequences if two economic variables that have been widely documented to contain unit roots, consumption and income, have no long-run relationship. Without a stable equilibrium relationship, the values of these two variables would diverge over time. Individuals would either have extremely high saving rates – when income is far above consumption, or become incredibly indebted.

¹⁴This is one version of the bootstrap and is appropriate for homoskedastic data. If the data are heteroskedastic, some form of block bootstrap is needed.

These two scenarios are implausible, and so there must be some long-run (or equilibrium) relationship between consumption and income. Similarly, consider the relationship between the spot and future price of oil. Standard finance theory dictates that future's price, f_t , is a conditionally unbiased estimate of the spot price in period $t + 1$, s_{t+1} ($E_t[s_{t+1}] = f_t$, assuming various costs such as the risk-free rate and storage are 0). Additionally, today's spot price is also an unbiased estimate of tomorrow's spot price ($E_t[s_{t+1}] = s_t$). However, both the spot and future price contain unit roots. Combining these two identities reveals a cointegrating relationship: $s_t - f_t$ should be stationary even if the spot and future prices contain unit roots.¹⁵

In stationary time-series, whether scalar or when the multiple processes are linked through a VAR, the process is self-equilibrating; given enough time, a process reverts to its unconditional mean. In a VAR, both the individual series and linear combinations of the series are stationary. The behavior of cointegrated processes is meaningfully different. Each component of a cointegrated process contains a unit root, and so has shocks with a permanent impact. However, when combined with another series, a cointegrated pair revert towards one another. A cointegrated pair is mean reverting to a stochastic trend (a unit root process), rather than to fixed value.

Cointegration and error correction provide a set of tools to analyze long-run relationships and short-term deviations from the equilibria. Cointegrated time-series exhibit temporary deviations from a long-run trend but are ultimately mean reverting to this trend. The Vector Error Correction Model (VECM) explicitly includes the deviation from the long-run relationship when modeling the short-term dynamics of the time series to push the components towards their long-run relationship.

5.8.1 Definition

Recall that a first-order integrated process is not stationary in levels but is stationary in differences. When this is the case, y_t is $I(1)$ and $\Delta y_t = y_t - y_{t-1}$ is $I(0)$. Cointegration builds on this structure by defining relationships across series which transform multiple $I(1)$ series into $I(0)$ series without using time-series differences.

Definition 5.13 (Bivariate Cointegration). Let $\{x_t\}$ and $\{y_t\}$ be two $I(1)$ series. These series are cointegrated if there exists a vector β with both elements non-zero such that

$$\beta' [x_t \ y_t]' = \beta_1 x_t - \beta_2 y_t \sim I(0) \quad (5.28)$$

This definition states that there exists a nontrivial linear combination of x_t and y_t that is stationary. This feature – a stable relationship between the two series, is a powerful tool in the analysis of nonstationary data. When treated individually, the data are extremely persistent; however, there is a well-behaved linear combination with transitory shocks that is stationary. Moreover, in many cases, this relationship takes a meaningful form such as $y_t - x_t$.

Cointegrating relationships are only defined up to a non-zero constant. For example if $x_t - \beta y_t$ is a cointegrating relationship, then $2x_t - 2\beta y_t = 2(x_t - \beta y_t)$ is also a cointegrating relationship. The standard practice is to normalize the vector on one of the variables so that its coefficient is unity. For example, if $\beta_1 x_t - \beta_2 y_t$ is a cointegrating relationship, the two normalized versions are $x_t - \beta_2/\beta_1 y_t = x_t - \tilde{\beta} y_t$ and $y_t - \beta_1/\beta_2 x_t = y_t - \tilde{\beta} x_t$.

The complete definition in the general case is similar, albeit slightly more intimidating.

¹⁵This assumes the horizon is short.

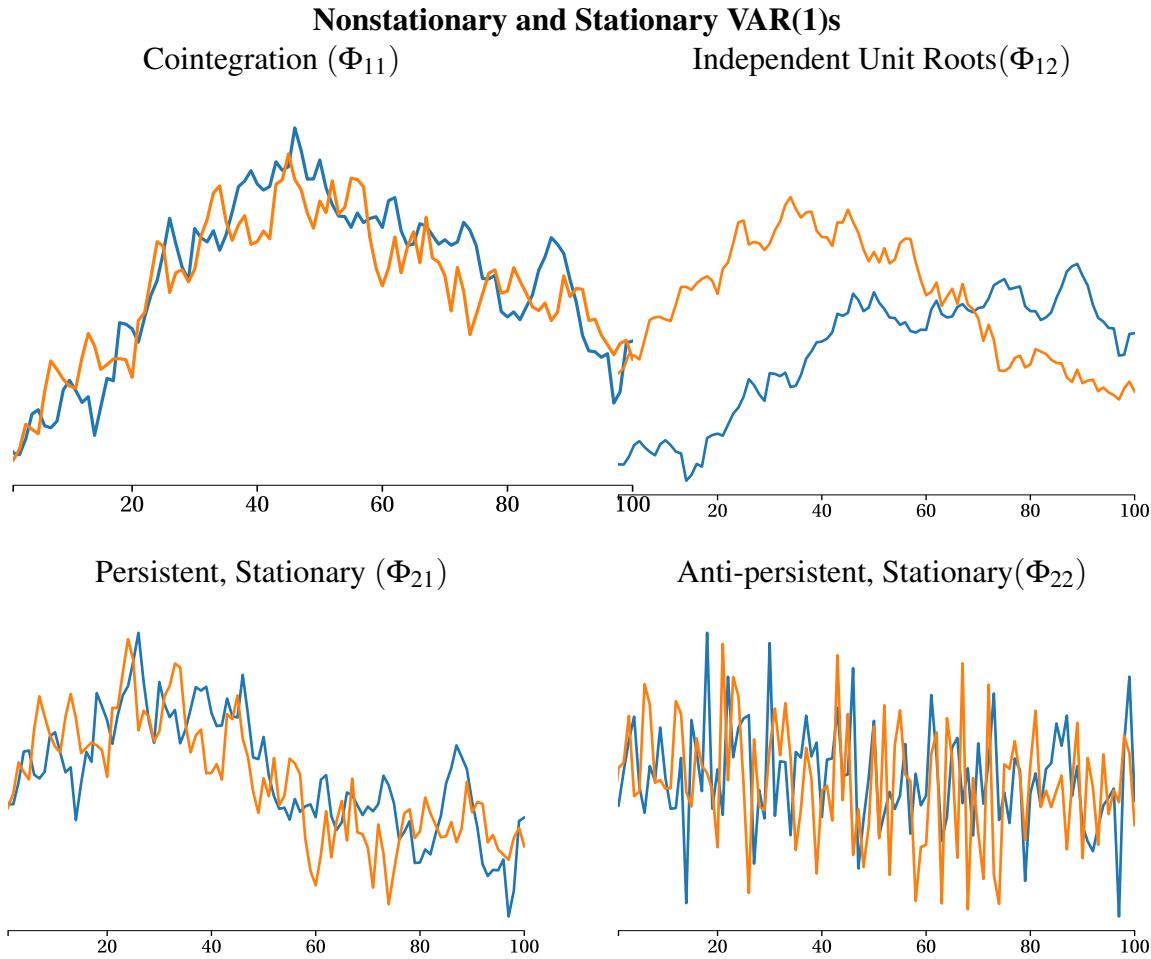


Figure 5.5: A plot of four time-series that all begin at the same initial value and use the same shocks. All data are generated by $\mathbf{y}_t = \Phi_{ij}\mathbf{y}_{t-1} + \varepsilon_t$ where Φ_{ij} varies across the panels.

Definition 5.14 (Cointegration). A set of k variables \mathbf{y}_t are cointegrated if at least two series are $I(1)$ and there exists a non-zero, reduced rank k by k matrix π such that

$$\pi\mathbf{y}_t \sim I(0). \quad (5.29)$$

The non-zero requirement is obvious: if $\pi = \mathbf{0}$ then $\pi\mathbf{y}_t = \mathbf{0}$ and this time series is trivially $I(0)$. The second requirement that π is reduced rank is not. This technical requirement is necessary since whenever π is full rank and $\pi\mathbf{y}_t \sim I(0)$, the series must be the case that \mathbf{y}_t is also $I(0)$. However, for variables to be *cointegrated*, they must be *integrated*. If the matrix is full rank, the common unit roots cannot cancel, and $\pi\mathbf{y}_t$ must have the same order of integration as \mathbf{y} . Finally, the requirement that at least two of the series are $I(1)$ rules out the degenerate case where all components of \mathbf{y}_t are $I(0)$, and allows \mathbf{y}_t to contain both $I(0)$ and $I(1)$ random variables. If \mathbf{y}_t contains both $I(0)$ and $I(1)$ random variables, then the long-run relationship only depends on the $I(1)$ random variable.

For example, suppose the components of $\mathbf{y}_t = [y_{1t}, y_{2t}]'$ are cointegrated so that $y_{1t} - \beta y_{2t}$ is sta-

tionary. One choice for π is

$$\pi = \begin{bmatrix} 1 & -\beta \\ 1 & -\beta \end{bmatrix}$$

To begin developing an understanding of cointegration, examine the plots in Figure 5.5. These four plots show two nonstationary processes and two stationary processes *all initialized at the same value and using the same shocks*. These plots contain simulated data from VAR(1) processes with different parameters, Φ_{ij} .

$$\mathbf{y}_t = \Phi_{ij}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t,$$

$$\Phi_{11} = \begin{bmatrix} .8 & .2 \\ .2 & .8 \end{bmatrix}, \quad \Phi_{12} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$\lambda_i = 1, 0.6 \quad \lambda_i = 1, 1$$

$$\Phi_{21} = \begin{bmatrix} .7 & .2 \\ .2 & .7 \end{bmatrix}, \quad \Phi_{22} = \begin{bmatrix} -.3 & .3 \\ .1 & -.2 \end{bmatrix},$$

$$\lambda_i = 0.9, 0.5 \quad \lambda_i = -0.43, -0.06$$

where λ_i are the eigenvalues of the parameter matrices. The nonstationary processes both have unit eigenvalues. The eigenvalues in the stationary processes are all less than 1 (in absolute value). The cointegrated process has a single unit eigenvalue while the independent unit root process has two. In a VAR(1), the number of unit eigenvalues plays a crucial role in cointegration and higher dimension cointegrated systems may contain between 1 and $k - 1$ unit eigenvalues. The number of unit eigenvalues shows the count of the unit root “drivers” in the system of equations.¹⁶ The picture presents evidence of the most significant challenge in cointegration analysis: it can be challenging to tell when two series are cointegrated, a feature in common with unit root testing of a single time series.

5.8.2 Vector Error Correction Models (VECM)

The Granger representation theorem provides a key insight into cointegrating relationships. Granger demonstrated that if a system is cointegrated then there exists a vector error correction model with a reduced rank coefficient matrix and if there is a VECM with a reduced rank coefficient matrix then the system must be cointegrated. A VECM describes the short-term deviations from the long-run trend (a stochastic trend/unit root). The simplest VECM is

$$\begin{bmatrix} \Delta x_t \\ \Delta y_t \end{bmatrix} = \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_{1,t} \\ \boldsymbol{\varepsilon}_{2,t} \end{bmatrix} \quad (5.30)$$

which states that changes in x_t and y_t are related to the levels of x_t and y_t through the cointegrating matrix (π). However, since x_t and y_t are cointegrated, there exists β such that $x_t - \beta y_t = [1 \ -\beta] [x_t \ y_t] \sim I(0)$. Substituting this value into this equation, equation 5.30 is equivalently expressed as

$$\begin{bmatrix} \Delta x_t \\ \Delta y_t \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \begin{bmatrix} 1 & -\beta \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_{1,t} \\ \boldsymbol{\varepsilon}_{2,t} \end{bmatrix}. \quad (5.31)$$

The short-run dynamics evolve according to

¹⁶In higher order VAR models, the eigenvalues must be computed from the companion form.

$$\Delta x_t = \alpha_1(x_{t-1} - \beta y_{t-1}) + \varepsilon_{1,t} \quad (5.32)$$

and

$$\Delta y_t = \alpha_2(x_{t-1} - \beta y_{t-1}) + \varepsilon_{2,t}. \quad (5.33)$$

The important elements of this VECM can be clearly labeled: $x_{t-1} - \beta y_{t-1}$ is the deviation from the long-run trend (also known as the equilibrium correction term) and α_1 and α_2 are the speed of adjustment parameters. VECMs impose one restriction of the α s: they cannot both be 0 (if they were, π would also be $\mathbf{0}$). In its general form, an VECM can be augmented to allow past short-run deviations to also influence present short-run deviations and to include deterministic trends. In vector form, an VECM(P) evolves according to

$$\Delta \mathbf{y}_t = \delta_0 + \pi \mathbf{y}_{t-1} + \pi_1 \Delta \mathbf{y}_{t-1} + \pi_2 \Delta \mathbf{y}_{t-2} + \dots + \pi_P \Delta \mathbf{y}_{t-P} + \varepsilon_t$$

where $\pi \mathbf{y}_{t-1} = \alpha \beta' \mathbf{y}_t$ captures the cointegrating relationship, δ_0 represents a linear time trend in the original data (levels) and $\pi_j \Delta \mathbf{y}_{t-j}$, $j = 1, 2, \dots, P$ capture short-run dynamics around the stochastic trend.

5.8.2.1 The Mechanics of the VECM

Any cointegrated VAR can be transformed into an VECM. Consider a simple cointegrated bivariate VAR(1)

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} .8 & .2 \\ .2 & .8 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}$$

To transform this VAR to an VECM, begin by subtracting $[x_{t-1} \ y_{t-1}]'$ from both sides

$$\begin{aligned} \begin{bmatrix} x_t \\ y_t \end{bmatrix} - \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} &= \begin{bmatrix} .8 & .2 \\ .2 & .8 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} - \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} \quad (5.34) \\ \begin{bmatrix} \Delta x_t \\ \Delta y_t \end{bmatrix} &= \left(\begin{bmatrix} .8 & .2 \\ .2 & .8 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} \\ \begin{bmatrix} \Delta x_t \\ \Delta y_t \end{bmatrix} &= \begin{bmatrix} -.2 & .2 \\ .2 & -.2 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} \\ \begin{bmatrix} \Delta x_t \\ \Delta y_t \end{bmatrix} &= \begin{bmatrix} -.2 \\ .2 \end{bmatrix} [1 \ -1] \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} \end{aligned}$$

In this example, the speed of adjustment parameters are $-.2$ for Δx_t and $.2$ for Δy_t and the normalized (on x_t) cointegrating relationship is $[1 \ -1]$.

In the general multivariate case, a cointegrated VAR(P) can be turned into an VECM by recursive substitution. Consider a cointegrated VAR(3),

$$\mathbf{y}_t = \Phi_1 \mathbf{y}_{t-1} + \Phi_2 \mathbf{y}_{t-2} + \Phi_3 \mathbf{y}_{t-3} + \varepsilon_t$$

This system is cointegrated if at least one but fewer than k eigenvalues of $\pi = \Phi_1 + \Phi_2 + \Phi_3 - \mathbf{I}_k$ are not zero. To begin the transformation, add and subtract $\Phi_3 \mathbf{y}_{t-2}$ to the right side

$$\begin{aligned}
\mathbf{y}_t &= \Phi_1 \mathbf{y}_{t-1} + \Phi_2 \mathbf{y}_{t-2} + \Phi_3 \mathbf{y}_{t-2} - \Phi_3 \mathbf{y}_{t-2} + \Phi_3 \mathbf{y}_{t-3} + \boldsymbol{\varepsilon}_t \\
&= \Phi_1 \mathbf{y}_{t-1} + \Phi_2 \mathbf{y}_{t-2} + \Phi_3 \mathbf{y}_{t-2} - \Phi_3 \Delta \mathbf{y}_{t-2} + \boldsymbol{\varepsilon}_t \\
&= \Phi_1 \mathbf{y}_{t-1} + (\Phi_2 + \Phi_3) \mathbf{y}_{t-2} - \Phi_3 \Delta \mathbf{y}_{t-2} + \boldsymbol{\varepsilon}_t.
\end{aligned}$$

Next, add and subtract $(\Phi_2 + \Phi_3) \mathbf{y}_{t-1}$ to the right-hand side,

$$\begin{aligned}
\mathbf{y}_t &= \Phi_1 \mathbf{y}_{t-1} + (\Phi_2 + \Phi_3) \mathbf{y}_{t-1} - (\Phi_2 + \Phi_3) \mathbf{y}_{t-1} + (\Phi_2 + \Phi_3) \mathbf{y}_{t-2} - \Phi_3 \Delta \mathbf{y}_{t-2} + \boldsymbol{\varepsilon}_t \\
&= \Phi_1 \mathbf{y}_{t-1} + (\Phi_2 + \Phi_3) \mathbf{y}_{t-1} - (\Phi_2 + \Phi_3) \Delta \mathbf{y}_{t-1} - \Phi_3 \Delta \mathbf{y}_{t-2} + \boldsymbol{\varepsilon}_t \\
&= (\Phi_1 + \Phi_2 + \Phi_3) \mathbf{y}_{t-1} - (\Phi_2 + \Phi_3) \Delta \mathbf{y}_{t-1} - \Phi_3 \Delta \mathbf{y}_{t-2} + \boldsymbol{\varepsilon}_t.
\end{aligned}$$

Finally, subtract \mathbf{y}_{t-1} from both sides,

$$\begin{aligned}
\mathbf{y}_t - \mathbf{y}_{t-1} &= (\Phi_1 + \Phi_2 + \Phi_3) \mathbf{y}_{t-1} - \mathbf{y}_{t-1} - (\Phi_2 + \Phi_3) \Delta \mathbf{y}_{t-1} - \Phi_3 \Delta \mathbf{y}_{t-2} + \boldsymbol{\varepsilon}_t \\
\Delta \mathbf{y}_t &= (\Phi_1 + \Phi_2 + \Phi_3 - \mathbf{I}_k) \mathbf{y}_{t-1} - (\Phi_2 + \Phi_3) \Delta \mathbf{y}_{t-1} - \Phi_3 \Delta \mathbf{y}_{t-2} + \boldsymbol{\varepsilon}_t.
\end{aligned}$$

The final step is to relabel the equation in terms of π notation,

$$\begin{aligned}
\mathbf{y}_t - \mathbf{y}_{t-1} &= (\Phi_1 + \Phi_2 + \Phi_3 - \mathbf{I}_k) \mathbf{y}_{t-1} - (\Phi_2 + \Phi_3) \Delta \mathbf{y}_{t-1} - \Phi_3 \Delta \mathbf{y}_{t-2} + \boldsymbol{\varepsilon}_t \quad (5.35) \\
\Delta \mathbf{y}_t &= \pi \mathbf{y}_{t-1} + \pi_1 \Delta \mathbf{y}_{t-1} + \pi_2 \Delta \mathbf{y}_{t-2} + \boldsymbol{\varepsilon}_t.
\end{aligned}$$

which is equivalent to

$$\Delta \mathbf{y}_t = \alpha \beta' \mathbf{y}_{t-1} + \pi_1 \Delta \mathbf{y}_{t-1} + \pi_2 \Delta \mathbf{y}_{t-2} + \boldsymbol{\varepsilon}_t. \quad (5.36)$$

where α contains the speed of adjustment parameters, and β contains the cointegrating vectors. This recursion can be used to transform any VAR(P), whether cointegrated or not,

$$\mathbf{y}_{t-1} = \Phi_1 \mathbf{y}_{t-1} + \Phi_2 \mathbf{y}_{t-2} + \dots + \Phi_P \mathbf{y}_{t-P} + \boldsymbol{\varepsilon}_t$$

into its VECM from

$$\Delta \mathbf{y}_t = \pi \mathbf{y}_{t-1} + \pi_1 \Delta \mathbf{y}_{t-1} + \pi_2 \Delta \mathbf{y}_{t-2} + \dots + \pi_{P-1} \Delta \mathbf{y}_{t-P+1} + \boldsymbol{\varepsilon}_t$$

using the identities $\pi = -\mathbf{I}_k + \sum_{i=1}^P \Phi_i$ and $\pi_p = -\sum_{i=p+1}^P \Phi_i$.¹⁷

¹⁷Stationary VAR(P) models can be written as VECM with one important difference. When $\{\mathbf{y}_t\}$ is covariance stationary, then π must have rank k . In cointegrated VAR models, the coefficient π in the VECM always has rank between 1 and $k-1$. If π has rank 0, then the VAR(P) contains k distinct unit roots and it is not possible to construct a linear combination that is $I(0)$.

5.8.2.2 Cointegrating Vectors

The key to understanding cointegration in systems with three or more variables is to note that the matrix which governs the cointegrating relationship, π , can always be decomposed into two matrices,

$$\pi = \alpha\beta'$$

where α and β are both k by r matrices where r is the number of cointegrating relationships. For example, suppose the parameter matrix in an VECM is

$$\pi = \begin{bmatrix} 0.3 & 0.2 & -0.36 \\ 0.2 & 0.5 & -0.35 \\ -0.3 & -0.3 & 0.39 \end{bmatrix}$$

The eigenvalues of this matrix are .9758, .2142 and 0. The 0 eigenvalue of π indicates there are two cointegrating relationships since the number of cointegrating relationships is $\text{rank}(\pi)$. Since there are two cointegrating relationships, β can be normalized to be

$$\beta = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \beta_1 & \beta_2 \end{bmatrix}$$

and α has 6 unknown parameters. $\alpha\beta'$ combine to produce

$$\pi = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{11}\beta_1 + \alpha_{12}\beta_2 \\ \alpha_{21} & \alpha_{22} & \alpha_{21}\beta_1 + \alpha_{22}\beta_2 \\ \alpha_{31} & \alpha_{32} & \alpha_{31}\beta_1 + \alpha_{32}\beta_2 \end{bmatrix},$$

and α can be determined using the left block of π . Once α is known, any two of the three remaining elements can be used to solve of β_1 and β_2 . Appendix A contains a detailed illustration of the steps used to find the speed of adjustment coefficients and the cointegrating vectors in trivariate cointegrated VARs.

5.8.3 Rank and the number of unit roots

The rank of π is the same as the number of distinct cointegrating vectors. Decomposing $\pi = \alpha\beta'$ shows that if π has rank r , then α and β must both have r linearly independent columns. α contains the speed of adjustment parameters, and β contains the cointegrating vectors. There are r cointegrating vectors, and so the system contains $m = k - r$ distinct unit roots. This relationship holds since when there are k variables and m distinct unit roots, it is always possible to find r distinct linear combinations eliminate the unit roots and so are stationary.

Consider a trivariate cointegrated system driven by either one or two unit roots. Denote the underlying unit root processes as $w_{1,t}$ and $w_{2,t}$. When there is a single unit root driving all three variables, the system can be expressed

$$\begin{aligned} y_{1,t} &= \kappa_1 w_{1,t} + \varepsilon_{1,t} \\ y_{2,t} &= \kappa_2 w_{1,t} + \varepsilon_{2,t} \\ y_{3,t} &= \kappa_3 w_{1,t} + \varepsilon_{3,t} \end{aligned}$$

where $\varepsilon_{j,t}$ is a covariance stationary error (or $I(0)$, but not necessarily white noise).

In this system there are two linearly independent cointegrating vectors. First consider normalizing the coefficient on $y_{1,t}$ to be 1 and so in the equilibrium relationship $y_{1,t} - \beta_1 y_{2,t} - \beta_2 y_{3,t}$ must satisfy

$$\kappa_1 = \beta_1 \kappa_2 + \beta_2 \kappa_3.$$

This equality ensures that the unit roots are not present in the difference. This equation does not have a unique solution since there are two unknown parameters. One solution is to further restrict $\beta_1 = 0$ so that the unique solution is $\beta_2 = \kappa_1 / \kappa_3$ and an equilibrium relationship is $y_{1,t} - (\kappa_1 / \kappa_3) y_{3,t}$. This alternative normalization produces a cointegrating vector since

$$y_{1,t} - \frac{\kappa_1}{\kappa_3} y_{3,t} = \kappa_1 w_{1,t} + \varepsilon_{1,t} - \frac{\kappa_1}{\kappa_3} \kappa_3 w_{1,t} - \frac{\kappa_1}{\kappa_3} \varepsilon_{3,t} = \varepsilon_{1,t} - \frac{\kappa_1}{\kappa_3} \varepsilon_{3,t}$$

Alternatively one could normalize the coefficient on $y_{2,t}$ and so the equilibrium relationship $y_{2,t} - \beta_1 y_{1,t} - \beta_2 y_{3,t}$ would require

$$\kappa_2 = \beta_1 \kappa_1 + \beta_2 \kappa_3.$$

This equation is also not identified since there are two unknowns and one equation. To solve assume $\beta_1 = 0$ and so the solution is $\beta_2 = \kappa_2 / \kappa_3$, which is a cointegrating relationship since

$$y_{2,t} - \frac{\kappa_2}{\kappa_3} y_{3,t} = \kappa_2 w_{1,t} + \varepsilon_{2,t} - \frac{\kappa_2}{\kappa_3} \kappa_3 w_{1,t} - \frac{\kappa_2}{\kappa_3} \varepsilon_{3,t} = \varepsilon_{2,t} - \frac{\kappa_2}{\kappa_3} \varepsilon_{3,t}$$

These solutions are the only two needed since any other definition of the equilibrium must be a linear combination of these. The redundant equilibrium is constructed by normalizing on $y_{1,t}$ to define an equilibrium of the form $y_{1,t} - \beta_1 y_{2,t} - \beta_2 y_{3,t}$. Imposing $\beta_3 = 0$ to identify the solution, $\beta_1 = \kappa_1 / \kappa_2$ which produces the equilibrium condition

$$y_{1,t} - \frac{\kappa_1}{\kappa_2} y_{2,t}.$$

This equilibrium is already implied by the first two,

$$y_{1,t} - \frac{\kappa_1}{\kappa_3} y_{3,t} \text{ and } y_{2,t} - \frac{\kappa_2}{\kappa_3} y_{3,t}$$

and can be seen to be redundant since

$$y_{1,t} - \frac{\kappa_1}{\kappa_2} y_{2,t} = \left(y_{1,t} - \frac{\kappa_1}{\kappa_3} y_{3,t} \right) - \frac{\kappa_1}{\kappa_2} \left(y_{2,t} - \frac{\kappa_2}{\kappa_3} y_{3,t} \right)$$

In this system of three variables and one common unit root the set of cointegrating vectors can be expressed as

$$\beta = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{\kappa_1}{\kappa_3} & \frac{\kappa_2}{\kappa_3} \end{bmatrix}.$$

When a system has only one unit root and three series, there are two non-redundant linear combinations of the underlying variables which are stationary. In a complete system with k variables and a single unit root, there are $k - 1$ non-redundant linear combinations that are stationary.

Next consider a trivariate system driven by two unit roots,

$$\begin{aligned}y_{1,t} &= \kappa_{11}w_{1,t} + \kappa_{12}w_{2,t} + \varepsilon_{1,t} \\y_{2,t} &= \kappa_{21}w_{1,t} + \kappa_{22}w_{2,t} + \varepsilon_{2,t} \\y_{3,t} &= \kappa_{31}w_{1,t} + \kappa_{32}w_{2,t} + \varepsilon_{3,t}\end{aligned}$$

where the errors $\varepsilon_{j,t}$ are again covariance stationary but not necessarily white noise. If the coefficient on $y_{1,t}$ is normalized to 1, then if the weights in the equilibrium condition, $y_{1,t} - \beta_1 y_{2,t} - \beta_2 y_{3,t}$, satisfy

$$\begin{aligned}\kappa_{11} &= \beta_1 \kappa_{21} + \beta_2 \kappa_{31} \\ \kappa_{12} &= \beta_1 \kappa_{22} + \beta_2 \kappa_{32}\end{aligned}$$

to order to eliminate both unit roots. This system of two equations in two unknowns has the solution

$$\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \kappa_{21} & \kappa_{31} \\ \kappa_{22} & \kappa_{32} \end{bmatrix}^{-1} \begin{bmatrix} \kappa_{11} \\ \kappa_{12} \end{bmatrix}.$$

This solution is unique (up to the initial normalization), and there are no other cointegrating vectors so that

$$\beta = \begin{bmatrix} 1 \\ \frac{\kappa_{11}\kappa_{32} - \kappa_{12}\kappa_{22}}{\kappa_{21}\kappa_{32} - \kappa_{22}\kappa_{31}} \\ \frac{\kappa_{12}\kappa_{21} - \kappa_{11}\kappa_{31}}{\kappa_{21}\kappa_{32} - \kappa_{22}\kappa_{31}} \end{bmatrix}$$

This line of reasoning extends to k -variate systems driven by m unit roots. One set of r cointegrating vectors is constructed by normalizing the first r elements of \mathbf{y} one at a time. In the general case

$$\mathbf{y}_t = \mathbf{K}\mathbf{w}_t + \varepsilon_t$$

where \mathbf{K} is a k by m matrix, \mathbf{w}_t an m by 1 set of unit root processes, and ε_t is a k by 1 vector of covariance stationary errors. Normalizing on the first r variables, the cointegrating vectors in this system are

$$\beta = \begin{bmatrix} \mathbf{I}_r \\ \tilde{\beta} \end{bmatrix} \tag{5.37}$$

where \mathbf{I}_r is an r -dimensional identity matrix. $\tilde{\beta}$ is a m by r matrix of loadings,

$$\tilde{\beta} = \mathbf{K}_2^{-1} \mathbf{K}'_1, \tag{5.38}$$

where \mathbf{K}_1 is the first r rows of \mathbf{K} (r by m) and \mathbf{K}_2 is the bottom m rows of \mathbf{K} (m by m). In the trivariate example driven by one unit root,

$$\mathbf{K}_1 = \begin{bmatrix} \kappa_1 \\ \kappa_2 \end{bmatrix} \text{ and } \mathbf{K}_2 = \kappa_3$$

and in the trivariate system driven by two unit roots,

$$\mathbf{K}_1 = [\kappa_{11} \ \kappa_{12}] \text{ and } \mathbf{K}_2 = \begin{bmatrix} \kappa_{21} & \kappa_{22} \\ \kappa_{31} & \kappa_{32} \end{bmatrix}.$$

Applying eqs. (5.37) and (5.38) produces the previously derived set of cointegrating vectors. Note that when $r = 0$ then the system contains k unit roots and so is not cointegrated (in general) since the system would have three equations and only two unknowns. Similarly when $r = k$ there are no unit roots and any linear combination is stationary.

5.8.3.1 Relationship to Common Features and common trends

Cointegration is a particular case of a broader concept known as common features. In the case of cointegration, both series have a common stochastic trend (or common unit root). Other examples of common features include common heteroskedasticity, defined as x_t and y_t are heteroskedastic but there exists a combination, $x_t - \beta y_t$, which is not, common nonlinearities which are defined analogously (replacing heteroskedasticity with nonlinearity), and cobreaks, where two series both contain structural breaks but $x_t - \beta y_t$ does now. Incorporating common features often produces simpler models than leaving them unmodeled.

5.8.4 Testing

Testing for cointegration, like testing for a unit root in a single series, is complicated. Two methods are presented, the original Engle-Granger 2-step procedure and the more sophisticated Johansen methodology. The Engle-Granger method is generally only applicable if there are two variables, if the system contains exactly one cointegrating relationship, or if the cointegration vector is known (e.g., an accounting identity where the left-hand side has to add up to the right-hand side). The Johansen methodology is substantially more general and can be used to examine complex systems with many variables and multiple cointegrating relationships.

5.8.4.1 Johansen Methodology

The Johansen methodology is the dominant technique used to determine whether a system of $I(1)$ variables is cointegrated and if so, to determine the number of cointegrating relationships. Recall that one of the requirements for a set of integrated variables to be cointegrated is that π has reduced rank,

$$\Delta \mathbf{y}_t = \pi \mathbf{y}_{t-1} + \pi_1 \Delta \mathbf{y}_{t-1} + \dots + \pi_P \Delta \mathbf{y}_{t-P} \varepsilon_t,$$

and the number of non-zero eigenvalues of π is between 1 and $k - 1$. If the number of non-zero eigenvalues is k , the system is stationary. If no non-zero eigenvalues are present, then the system contains k unit roots, is not cointegrated and it is not possible to define a long-run relationship. The Johansen framework for cointegration analysis uses the magnitude of the eigenvalues of $\hat{\pi}$ to test for cointegration. The Johansen methodology also allows the number of cointegrating relationships to be determined from the data directly, a key feature missing from the Engle-Granger two-step procedure.

The Johansen methodology makes use of two statistics, the trace statistic (λ_{trace}) and the maximum eigenvalue statistic (λ_{max}). Both statistics test functions of the estimated eigenvalues of π but have different null and alternative hypotheses. The trace statistic tests the null that the number of cointegrating relationships is less than or equal to r against an alternative that the number is greater than r .

Define $\hat{\lambda}_i$, $i = 1, 2, \dots, k$ to be the complex modulus of the eigenvalues of $\hat{\pi}_1$ and let them be ordered such that $\lambda_1 > \lambda_2 > \dots > \lambda_k$.¹⁸ The trace statistic is defined

$$\lambda_{\text{trace}}(r) = -T \sum_{i=r+1}^k \ln(1 - \hat{\lambda}_i).$$

There are k trace statistics. The trace test is applied sequentially, and the number of cointegrating relationships is determined by proceeding through the test statistics until the null is not rejected. The first trace statistic, $\lambda_{\text{trace}}(0) = -T \sum_{i=1}^k \ln(1 - \hat{\lambda}_i)$, tests the null there are no cointegrating relationships (i.e., the system contains k unit roots) against an alternative that the number of cointegrating relationships is one or more. If there are no cointegrating relationships, then the true rank of π is 0, and each of the estimated eigenvalues should be close to zero. The test statistic $\lambda_{\text{trace}}(0) \approx 0$ since every unit root “driver” corresponds to a zero eigenvalue in π . When the series are cointegrated, π has one or more non-zero eigenvalues. If only one eigenvalue is non-zero, so that $\lambda_1 > 0$, then in large samples $\ln(1 - \hat{\lambda}_1) < 0$ and $\lambda_{\text{trace}}(0) \approx -T(1 - \lambda_1)$, which becomes arbitrarily large as T grows.

Like unit root tests, cointegration tests have nonstandard distributions that depend on the included deterministic terms if any. Software packages return the appropriate critical values for the length of the time-series analyzed and included deterministic regressors if any.

The maximum eigenvalue test examines the null that the number of cointegrating relationships is r against the alternative that the number is $r + 1$. The maximum eigenvalue statistic is defined

$$\lambda_{\max}(r, r+1) = -T \ln(1 - \hat{\lambda}_{r+1})$$

Intuitively, if there are $r + 1$ cointegrating relationships, then the $r + 1^{\text{th}}$ ordered eigenvalue should be positive, $\ln(1 - \hat{\lambda}_{r+1}) < 0$, and the value of $\lambda_{\max}(r, r+1) \approx -T \ln(1 - \lambda_{r+1})$ should be large. On the other hand, if there are only r cointegrating relationships, the $r + 1^{\text{th}}$ eigenvalue is zero, its estimate should be close to zero, and so the statistic should be small. Again, the distribution is nonstandard, but statistical packages provide appropriate critical values for the number of observations and the included deterministic regressors.

The steps to implement the Johansen procedure are:

Step 1: Plot the data series being analyzed and perform univariate unit root testing. A set of variables can only be *cointegrated* if they are all *integrated*. If the series are trending, either linearly or quadratically, remember to include deterministic terms when estimating the VECM.

Step 2: The second stage is lag length selection. Select the lag length using one of the procedures outlined in the VAR lag length selection section (e.g., General-to-Specific or AIC). For example, to use the General-to-Specific approach, first select a maximum lag length L and then, starting with $l = L$, test l lags against $l - 1$ use a likelihood ratio test,

$$LR = (T - l \cdot k^2)(\ln |\Sigma_{l-1}| - \ln |\Sigma_l|) \sim \chi_k^2.$$

Repeat the test by decreasing the number of lags (l) until the LR rejects the null that the smaller model is equivalent to the larger model.

Step 3: Estimate the selected VECM,

¹⁸The complex modulus is defined as $|\lambda_i| = |a + bi| = \sqrt{a^2 + b^2}$.

$$\Delta \mathbf{y}_t = \pi_0 + \pi_1 \Delta \mathbf{y}_{t-1} + \dots + \pi_{P-1} \Delta \mathbf{y}_{t-P+1} + \varepsilon$$

and determine the rank of π where P is the lag length previously selected. If the levels of the series appear to be trending, then the model in differences should include a constant and

$$\Delta \mathbf{y}_t = \delta_0 + \pi_1 \Delta \mathbf{y}_{t-1} + \dots + \pi_{P-1} \Delta \mathbf{y}_{t-P+1} + \varepsilon$$

should be estimated. Using the λ_{trace} and λ_{\max} tests, determine the cointegrating rank of the system. It is important to check that the residuals are weakly correlated – so that there are no important omitted variables, the residuals are not excessively heteroskedastic, which affects the size and power of the procedure, and are approximately Gaussian.

Step 4: Analyze the normalized cointegrating vectors to determine whether these conform to implications of finance theory. Hypothesis tests on the cointegrating vector can also be performed to examine whether the long-run relationships conform to a particular theory.

Step 5: The final step of the procedure is to assess the adequacy of the model by plotting and analyzing the residuals. This step should be the final task in the analysis of any time-series data, not just the Johansen methodology. If the residuals do not resemble white noise, the model should be reconsidered. If the residuals are stationary but autocorrelated, more lags may be necessary. If the residuals are $I(1)$, the system may not be cointegrated.

Lag Length Selection

Tests of cointegration using the two test statistic, λ_{trace} and λ_{\max} , are sensitive to the lag length. The number of included lags must be sufficient to produce white noise residuals. The lag length is commonly chosen using an IC, and given the trade-off between a model that is too small – which leaves serial correlation in the model residuals – and too large, which produces noisier estimates of parameters but no serial correlation, a loose criterion like the AIC is preferred to a more strict one.

Trends

Nonstationary time series often contain time trends. Like the Augmented Dickey-Fuller test, Johansen's λ_{trace} and λ_{\max} tests are both sensitive to the choice of included trends. There are five different configurations of trends in the VECM,

$$\Delta \mathbf{y}_t = \delta_0 + \delta_1 t + \alpha' (\beta \mathbf{y}_{t-1} + \gamma_0 + \gamma_1 t) + \pi_1 \Delta \mathbf{y}_{t-1} + \dots + \pi_{P-1} \Delta \mathbf{y}_{t-P+1} + \varepsilon.$$

The five test configurations of the test are:

- no trends, $\delta_0 = \delta_1 = \gamma_0 = \gamma_1 = \mathbf{0}$;
- linear trend in \mathbf{y}_t , $\alpha' \beta \mathbf{y}_{t-1}$ is mean 0, $\delta_1 = \gamma_0 = \gamma_1 = \mathbf{0}$;
- linear trend in \mathbf{y}_t , non-zero mean $\alpha' \beta \mathbf{y}_{t-1}$, $\delta_1 = \gamma_1 = \mathbf{0}$;
- quadratic trend in \mathbf{y}_t , non-zero mean $\alpha' \beta \mathbf{y}_{t-1}$, $\gamma_1 = \mathbf{0}$; and
- quadratic trend in \mathbf{y}_t , linear trend in $\alpha' \beta \mathbf{y}_{t-1}$, no restrictions on the parameters.

The simplest specification sets all trends to be 0, so that

$$\Delta \mathbf{y}_t = \alpha' \beta \mathbf{y}_{t-1} + \pi_1 \Delta \mathbf{y}_{t-1} + \dots + \pi_{P-1} \Delta \mathbf{y}_{t-P+1} + \varepsilon.$$

The specification is only appropriate if the components of \mathbf{y}_t are not trending. When the component time series of \mathbf{y}_t have linear time trends, then

$$\Delta \mathbf{y}_t = \delta_0 + \alpha' (\beta \mathbf{y}_{t-1} + \gamma_0) + \pi_1 \Delta \mathbf{y}_{t-1} + \dots + \pi_{P-1} \Delta \mathbf{y}_{t-P+1} + \varepsilon$$

allows them to appear in two places. The two intercepts, δ_0 and $\alpha' \gamma_0$ play different roles. δ_0 allows for time trends in the component series since the left-hand-side has been differenced, so that a time-trend in the level becomes an intercept in the difference. γ_0 allows the cointegrating relationship to have a non-zero mean, which is practically important in many applications of cointegration. The model can be estimated assuming $\gamma_0 = \mathbf{0}$ so that

$$\Delta \mathbf{y}_t = \delta_0 + \alpha' \beta \mathbf{y}_{t-1} + \pi_1 \Delta \mathbf{y}_{t-1} + \dots + \pi_{P-1} \Delta \mathbf{y}_{t-P+1} + \varepsilon.$$

In this specification, the components are allowed to have unrestricted time trends but the cointegrating relationships are restricted to be mean zero. In practice, this requires that the growth rates of the component time series in \mathbf{y}_t are the same. The full set of time trends are included in the model, the \mathbf{y}_t is allowed to have a quadratic time trend (the difference has a linear time trend) and the cointegrating relationship,

$$\beta \mathbf{y}_{t-1} + \gamma_0 + \gamma_1 t$$

may also have a time trend. The specification with a time trend can be restricted so that $\gamma_1 = \mathbf{0}$ in which case the cointegrating relationships are allowed to have a mean different from $\mathbf{0}$ but not to be trending.

Additional trend components increase the critical values of the λ_{trace} and λ_{\max} test statistics, and so, all things equal, it is harder to reject the null. The principle behind selecting deterministic terms in the Johansen's framework is the same as when including deterministic terms in ADF tests – any deterministic that is present in the data must be included, and failing to include a required deterministic term prevents the null from being rejected even in large samples. Similarly, including more deterministic trends than required lowers the power of the test and so makes it more challenging to find cointegration when it is present. Deterministic trends should be eliminated using a general-to-specific search starting with the full set of terms, and eliminating any that are (jointly) insignificant.

5.8.4.2 Example: Consumption Aggregate Wealth

To illustrate cointegration and error correction, three series which have revived the CCAPM in recent years are examined (Lettau and Ludvigson, 2001a; Lettau and Ludvigson, 2001b). These three series are consumption (c), asset prices (a) and labor income (y). The data are made available by [Martin Lettau on his web site](#), and contain quarterly data from 1952:1 until 2017:3. These series are documented to be cointegrated in published papers, and the cointegrating error is related to expected future returns. When $c - \delta_0 - \beta_a a - \beta_y$ is positive, then consumption is above its long-run trend, and so asset returns are expected to be above average. When this error is negative, then c is relatively low compared to asset values and labor income, and so asset values are too high.

Trace Test			
Null	Alternative	λ_{trace}	Crit. Val.
$r = 0$	$r \geq 1$	19.06	29.80
$r = 1$	$r \geq 2$	8.68	15.49
$r = 2$	$r = 3$	2.03	3.84

Max Test			
Null	Alternative	λ_{\max}	Crit. Val.
$r = 0$	$r = 1$	10.39	21.13
$r = 1$	$r = 2$	6.64	14.26
$r = 2$	$r = 3$	2.03	3.84

Table 5.5: Results of testing using the Johansen methodology to the *cay* time series.

The Johansen methodology begins by examining the original data for unit roots. The results in Table 5.6 establish that all series have unit roots using ADF tests. The next step tests eigenvalues of π in the VECM

$$\Delta \mathbf{y}_t = \delta_0 + \pi(\mathbf{y}_{t-1} + \gamma_0) + \pi_1 \Delta \mathbf{y}_{t-1} + \pi_2 \Delta \mathbf{y}_{t-2} + \dots + \pi_P \Delta \mathbf{y}_{t-P} + \varepsilon_t.$$

using λ_{trace} and λ_{\max} tests. Table 5.5 contains the results of the two tests. These tests are applied sequentially. The first null hypothesis is not rejected for either test, which indicates that the π has rank 0, and so the system contains three distinct unit roots, and so the variables are not cointegrated.¹⁹

5.8.4.3 A Single Cointegrating Relationship: Engle-Granger Methodology

The Engle-Granger method exploits the defining characteristic of a cointegrated system with a single cointegrating relationship – if the time series are cointegrated, then a linear combination of the series can be constructed that is stationary. If they are not, then any linear combination remains $I(1)$. When there are two variables, the Engle-Granger methodology begins by specifying the cross-section regression

$$y_t = \beta x_t + \varepsilon_t$$

where $\hat{\beta}$ can be estimated using OLS. It may be necessary to include a constant,

$$y_t = \delta_0 + \beta x_t + \varepsilon_t$$

or a constant and time trend,

$$y_t = \delta_0 + \delta_1 t + \beta x_t + \varepsilon_t,$$

if the residuals from the simple cross-sectional regression are not mean 0 or trending. The model residuals, $\hat{\varepsilon}_t$, are constructed from the OLS estimates of the model coefficients and are tested for the

¹⁹The first null not rejected indicates the cointegrating rank of the system. If all null hypotheses are rejected, then the original system appears stationary, and a reanalysis of the $I(1)$ classification of the original data is warranted.

presence of a unit root. If x_t and y_t are both $I(1)$ and $\hat{\varepsilon}_t$ is $I(0)$, then the series are cointegrated. If the null that $\hat{\varepsilon}_t$ contains a unit root is not rejected, then the two series are no cointegrated since the difference did not eliminate the unit root. The procedure concludes by using $\hat{\varepsilon}_t$ to estimate the VECM to estimate parameters which may be of interest (e.g., the speed of convergence parameters).

Step 1: Begin by analyzing x_t and y_t in isolation to ensure that they are both integrated, plot the data, and perform ADF tests. Remember, variables can only be *cointegrated* if they are *integrated*.

Step 2: Estimate the long-run relationship by fitting

$$y_t = \delta_0 + \delta_1 t + \beta x_t + \varepsilon_t,$$

where the two deterministic terms are included only if necessary, using OLS and computing the estimated residuals $\{\hat{\varepsilon}_t\}$. Use an ADF test (or DF-GLS for more power) and test $H_0 : \gamma = 0$ against $H_1 : \gamma < 0$ in the regression

$$\Delta\hat{\varepsilon}_t = \gamma\hat{\varepsilon}_{t-1} + \psi_1\Delta\hat{\varepsilon}_{t-1} + \dots + \psi_p\Delta\hat{\varepsilon}_{t-P} + \eta_t.$$

Deterministic effects are removed in the cross-sectional regression, and so are not included in the ADF test. If the null is rejected and $\hat{\varepsilon}_t$ is stationary, then x_t and y_t appear to be cointegrated. Alternatively, if $\hat{\varepsilon}_t$ still contains a unit root, the series are not cointegrated.²⁰

Step 3: If a cointegrating relationship is found, specify and estimate the VECM

$$\begin{bmatrix} \Delta x_t \\ \Delta y_t \end{bmatrix} = \begin{bmatrix} \pi_{01} \\ \pi_{02} \end{bmatrix} + \begin{bmatrix} \alpha_1(y_{t-1} - \delta_0 - \delta_1 t - \beta x_{t-1}) \\ \alpha_2(y_{t-1} - \delta_0 - \delta_1 t - \beta x_{t-1}) \end{bmatrix} + \pi_1 \begin{bmatrix} \Delta x_{t-1} \\ \Delta y_{t-1} \end{bmatrix} + \dots + \pi_P \begin{bmatrix} \Delta x_{t-P} \\ \Delta y_{t-P} \end{bmatrix} + \begin{bmatrix} \eta_{1,t} \\ \eta_{2,t} \end{bmatrix}$$

Note that this specification is not linear in its parameters. Both equations have interactions between the α and β parameters and so OLS cannot be used. Engle and Granger noted that the terms involving β can be replaced with $\hat{\varepsilon}_{t-1} = (y_{t-1} - \hat{\beta}_1 - \hat{\beta}_2 x_{t-1})$,

$$\begin{bmatrix} \Delta x_t \\ \Delta y_t \end{bmatrix} = \begin{bmatrix} \pi_{01} \\ \pi_{02} \end{bmatrix} + \begin{bmatrix} \alpha_1\hat{\varepsilon}_{t-1} \\ \alpha_2\hat{\varepsilon}_{t-1} \end{bmatrix} + \pi_1 \begin{bmatrix} \Delta x_{t-1} \\ \Delta y_{t-1} \end{bmatrix} + \dots + \pi_P \begin{bmatrix} \Delta x_{t-P} \\ \Delta y_{t-P} \end{bmatrix} + \begin{bmatrix} \eta_{1,t} \\ \eta_{2,t} \end{bmatrix},$$

and so parameters of these specifications can be estimated using OLS. The substitution has no impact on the standard errors of the estimated parameters since the parameters of the cointegrating relationship are super-consistent (i.e., they converge faster than the standard \sqrt{T} rate).

Step 4: The final step is to assess the model adequacy and test hypotheses about α_1 and α_2 . Standard diagnostic checks including plotting the residuals and examining the ACF should be used to examine model adequacy. Impulse response functions for the short-run deviations can be examined to assess the effect of a shock on the deviation of the series from the long term trend.

Deterministic Regressors

The cross-sectional regression in the Engle-Granger methodology can be modified to accommodate three configurations of deterministic regressors. The simplest configuration has no deterministic terms so that the regression is

$$y_t = \beta x_t + \varepsilon_t.$$

²⁰The distribution of the ADF is different when testing cointegration than when testing for a unit root. Software packages report the correct value which depends on the number of variables in the cointegrating relationship and the deterministic terms if any.

Engle-Granger is only limited finding a single cointegrating relationship, which might exist between k variables, not just 2. In this case, the cross-sectional regression is

$$y_t = \beta' \mathbf{x}_t + \varepsilon_t$$

where \mathbf{x}_t is the $k - 1$ by 1 vector, and the cointegrating vector is $[1, -\beta']$. This generalized form can be further extended by altering the deterministic terms in the model. For example, it is common to include an intercept in the cross-sectional regression,

$$y_t = \delta_0 + \beta' \mathbf{x}_t + \varepsilon_t.$$

This structure allows the long-run relationship between y_t and \mathbf{x}_t to have a non-zero mean. The intercept should be included except where theory suggests that the cointegrating errors should be zero, e.g., in the relationship between spot and future prices or the long-run relationship between prices of the same security trading in different markets.

The cross-sectional regression can be further extended to include a time trend,

$$y_t = \delta_0 + \delta_1 t + \beta' \mathbf{x}_t + \varepsilon_t.$$

When the model includes a time-trend, the long-run relationship, y_t and \mathbf{x}_t , is assumed to be trending over time, so that $y_t - \delta_0 - \delta_1 t - \beta' \mathbf{x}_t$ is a mean-zero $I(0)$ process. This might occur if the growth rates of y_t and the components \mathbf{x}_t differ. It is much less common to include time-trends in the cointegrating relationship. Best practice is to only include δ_1 if there is some *a priori* reason to believe that the relationship has a time-trend and when $\hat{\delta}_1$ is statistically different from 0 when the cross-sectional regression is estimated. The cross-sectional regression can be compactly expressed as

$$y_t = \delta' \mathbf{d}_t + \beta' \mathbf{x}_t + \varepsilon_t$$

where \mathbf{d}_t is the vector of included deterministic regressors, i.e., one of [] (nothing), [1], or [1, t].

Dynamic OLS

The parameter estimators of the cointegrating vector estimated using a cross-sectional regression is not normally distributed in large samples. It is also not efficient since the $I(1)$ variables might have short-run dynamics. Dynamic OLS, a simple modification of the Engle-Granger regression, addresses both of these. It adds lags and *leads* of the differences on the right-hand-side variables to the cross-sectional regression. These extra terms effectively remove the short term dynamics in the right-hand-side variables. In a bivariate cointegrated relationship, the Dynamic OLS regression is

$$y_t = \delta' \mathbf{d}_t + \beta_1 x_t + \sum_{i=-P}^P \gamma_i \Delta x_{t-i} + \varepsilon_t$$

where \mathbf{d}_t is a vector of deterministic terms in the model. This regression is estimated using OLS, and the estimated cointegrating relationship is $y_t - \hat{\delta}' \mathbf{d}_t - \hat{\beta}_1 x_t$. If there are more than 1-right-hand-side variables, then the regression is

$$y_t = \delta' \mathbf{d}_t + \beta' \mathbf{x}_t + \sum_{i=-P}^P \gamma' \Delta \mathbf{x}_{t-i} + \varepsilon_t$$

Unit Root Tests			
Series	T-stat	P-val	ADF Lags
c	-1.198	0.674	5
a	-0.205	0.938	3
y	-2.302	0.171	0
$\hat{\epsilon}_t^c$	-2.706	0.383	1
$\hat{\epsilon}_t^a$	-2.573	0.455	0
$\hat{\epsilon}_t^y$	-2.679	0.398	1

Table 5.6: The top three lines contain the results of ADF tests for unit roots in the three components of cay : Consumption, Asset Prices and Aggregate Wealth. The final lines contain the results of unit root tests on the estimated residuals from the cross-sectional regressions. The variable in the superscript is the dependent variable in the Engle-Granger regression. The lags column reports the number of lags used in the ADF procedure, which is automatically selected using the AIC.

Comparing Engle-Granger and Dynamic OLS						
	Dependent Variable					
	c		a		y	
δ_0	-0.643	-0.640 (-6.896)	1.917	1.874 (7.784)	0.702	0.713 (5.312)
β_c			2.284	2.385 (6.994)	1.163	1.180 (18.521)
β_a	0.249	0.260 (6.187)			-0.214	-0.229 (-3.790)
β_y	0.785	0.773 (17.339)	-1.322	-1.421 (-4.024)		

Table 5.7: Each column reports estimates of the cointegrating relationship where the dependent variable varies across the three series. The parameter estimators in Engle-Granger regressions are not asymptotically normally distributed, and so t-stats are not reported. The t-stats reported for the estimates produced using Dynamic OLS are computed using the Newey-West covariance estimator with 14 lags.

where β , γ_i and \mathbf{x}_t are $k - 1$ by 1 vectors. The estimators of the cointegrating vector are asymptotically normally distributed, although the parameter covariance must be estimated using a long-run covariance estimator that accounts for dependence, e.g., Newey-West (see Section 5.9.2). The number of leads and lags to include in the model can be selected using an information criterion. In application in macrofinance, it is often chosen to capture 1 year of data, so either 4 (quarterly) or 12 (monthly).

5.8.4.4 Cointegration in Consumption, Asset Prices and Income

The Engle-Granger procedure begins by performing unit root tests on the individual series and examining the data. Table 5.6 and contain the results from ADF tests and Figure 5.6 plots the detrended series. The null of a unit root is not rejected in any of the three series, and all have time-detrended errors which appear to be nonstationary.

The next step is to specify the cointegrating regression

$$c_t = \delta_0 + \beta_a a_t + \beta_y y_t + \varepsilon_t$$

and to estimate the long-run relationship using OLS. The estimated cointegrating vector from is $[1 - 0.249 - 0.785]$, and corresponds to a long-run relationship of $\hat{\varepsilon}_t = c_t + .643 - 0.249a_t - 0.785y_t$. Finally, the residuals are tested for the presence of a unit root. The results of this test are labeled $\hat{\varepsilon}_t^c$ in Table 5.6 and indicate that the null is not rejected, and so the three series are not cointegrated. The Engle-Granger methodology agrees with the Johansen methodology that it is not possible to eliminate the unit roots from the three series using a single linear combination. It is also possible to normalize the coefficients on a or y by using these are the dependent variable. The final two lines in Table 5.6 contain results for these specifications. The results for the alternative agree with the finding for c , and the series do not appear to be cointegrated. The middle panel of Figure 5.6 plot the three residual series where each of the variables is used as the dependent. The residuals constructed from the regression when a or y are the dependent are multiplied by -1 so that the sign on c is always positive, and all three series are normalized to have unit variance (for comparability). The three residual series are very similar which indicates that the choice of the dependent variable has little impact on the estimates of the cointegrating relationship.

The VECM uses the residuals estimated using the cross-sectional regression, $\hat{\varepsilon}_t = c_t - \hat{\delta}_0 - \hat{\beta}_a a_t - \hat{\beta}_y y_t$.

$$\begin{bmatrix} \Delta c_t \\ \Delta a_t \\ \Delta y_t \end{bmatrix} = \begin{bmatrix} 0.003 \\ 0.004 \\ 0.003 \end{bmatrix} + \begin{bmatrix} -0.000 \\ 0.002 \\ 0.000 \end{bmatrix} \hat{\varepsilon}_{t-1} + \begin{bmatrix} 0.192 & 0.102 & 0.147 \\ (0.005) & (0.000) & (0.004) \\ 0.282 & 0.220 & -0.149 \\ (0.116) & (0.006) & (0.414) \\ 0.369 & 0.061 & -0.139 \\ (0.000) & (0.088) & (0.140) \end{bmatrix} \begin{bmatrix} \Delta c_{t-1} \\ \Delta a_{t-1} \\ \Delta y_{t-1} \end{bmatrix} + \eta_t$$

The coefficients on the lagged residual measure the speed of adjustment. The estimates are all close to 0 indicating that deviations from the equilibrium are highly persistent. Two of the speed of adjustment coefficients are not statistically different from zero, which indicates that three series are not well described as a cointegrated system. The lag length in the VECM is selected by minimizing the HQIC using up to 4 lags of the quarterly data.

Table 5.7 contains estimates of the parameters from the Engle-Granger cross-sectional regressions and the Dynamic OLS regressions. The DOLS estimates are asymptotically normal (if the series are cointegrated) and so standard errors, computed using the Newey-West covariance estimator, are reported for the coefficients. The bottom panel of Figure 5.6 plot the residual from the two estimators when c is the dependent variable. The leads and lags have little effect on the estimated cointegration vector, and so the two series are very similar.

5.8.5 Spurious Regression and Balance

When a regression is estimated using two related $I(1)$ variables, the cointegrating relationship dominates and the regression coefficients can be directly interpreted as the cointegrating vectors. However, when a model is estimated on two unrelated $I(1)$ variables, the regression estimator is no longer consistent. For example, let x_t and y_t be independent random walk processes.

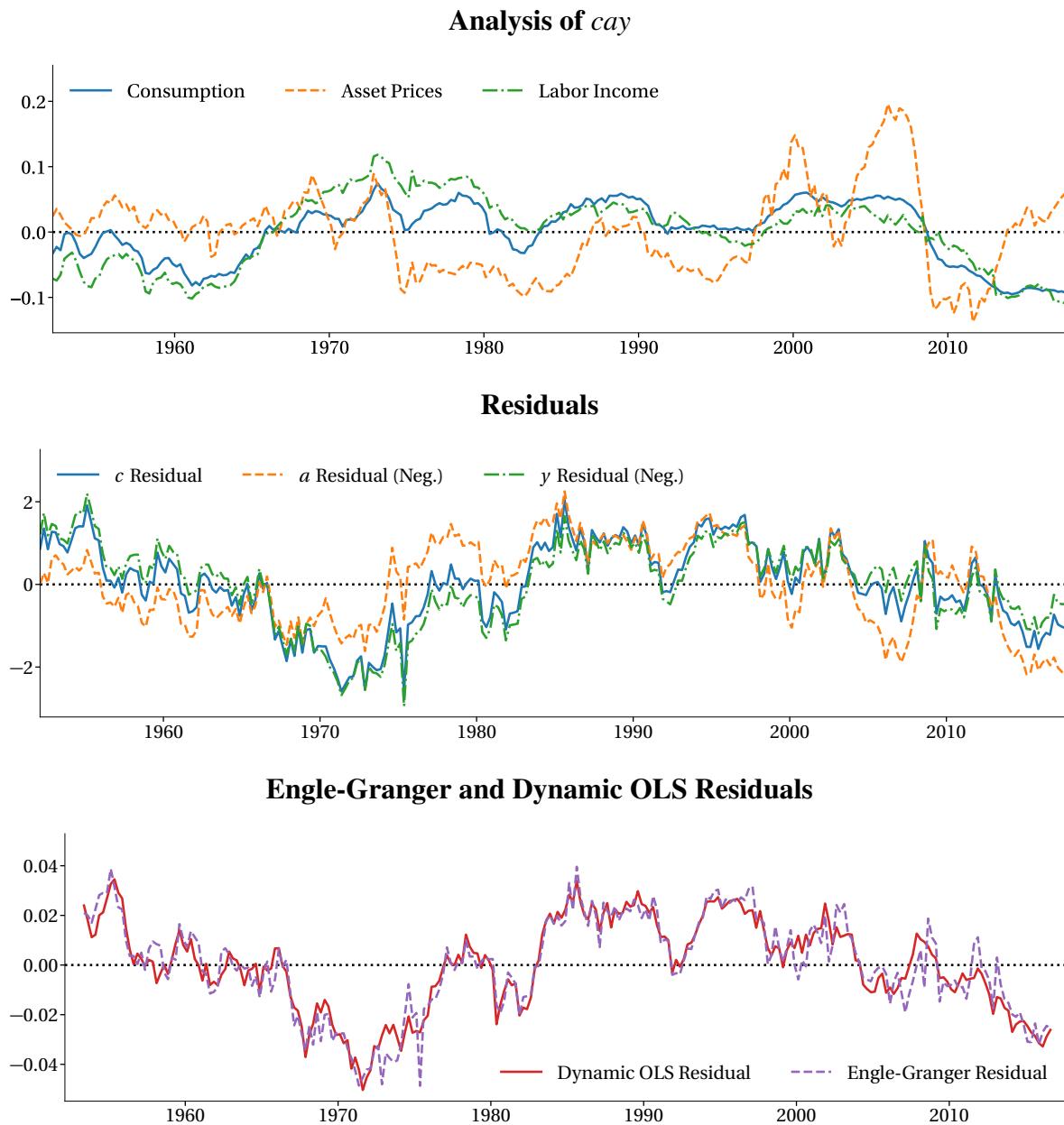


Figure 5.6: The top panel contains plots of detrended residuals from regressions of consumption, asset prices and labor income on a linear time trend. The middle panel contains a plot of residuals from the three specifications of the Engle-Granger regression where each of the three series is used as the dependent variable. The residuals are multiplied by -1 when a or y is the dependent variable so they the sign on c is always positive. The residuals are all normalized to have unit variance. The bottom panel plots the residuals computed using the Dynamic OLS estimates of the cointegrating relationship when c is the dependent variable and 4 leads and lags are used.

$$x_t = x_{t-1} + \eta_t$$

and

$$y_t = y_{t-1} + \nu_t$$

In the regression

$$x_t = \beta y_t + \varepsilon_t$$

$\hat{\beta}$ is not consistent for 0 despite the independence of x_t and y_t .

Models that include independent $I(1)$ processes are known as **spurious regressions**. When the regressions are spurious, the estimated $\hat{\beta}$ can take any value and typically have t -stats that indicate significance at conventional levels. The solution to this problem is simple: whenever regressing one $I(1)$ variable on another, always check to be sure that the regression residuals are $I(0)$ and not $I(1)$ – in other words, verify that the series are cointegrated. If the series are not cointegrated, it is not possible to estimate a meaningful long-run relationship between the two (or more) $I(1)$ random variables. Nonstationary time series that are not cointegrated can be differenced to be $I(0)$ and then modeled as a stationary VAR.

Balance is an important concept when data which contain both stationary and integrated data. An equation is balanced if all variables have the same order of integration. The usual case occurs when a stationary variable ($I(0)$) is related to one or more other stationary variables. It is illustrative to consider the four combinations:

- $I(0)$ on $I(0)$: The usual case. Standard asymptotic arguments apply. See section 5.9 for more issues in cross-section regression using time-series data.
- $I(1)$ on $I(0)$: This regression is unbalanced. An $I(0)$ variable can never explain the long-run variation in an $I(1)$ variable. The usual solution is to difference the $I(1)$ and then examine whether the short-run dynamics in the differenced $I(1)$, which are $I(0)$, can be explained by the $I(0)$.
- $I(1)$ on $I(1)$: One of two outcomes: cointegration or spurious regression.
- $I(0)$ on $I(1)$: This regression is unbalanced. An $I(1)$ variable can never explain the variation in an $I(0)$ variable, and unbalanced regressions are not useful tools for explaining economic phenomena. Unlike spurious regressions, the t -stat still has a standard asymptotic distribution although caution is needed since the CLT does not, in empirically relevant sample sizes, provide an accurate approximation to the finite sample distribution. Poor finite-sample approximations are common in applications where a stationary variable, e.g., returns on the market, is regressed on a highly persistent predictor (such as the default premium, dividend yield or price-to-earnings ratio).

5.9 Cross-sectional Regression with Time-series Data

Cross-sectional regressions are commonly estimated using data that occur sequentially, e.g., the CAP-M and related models. Chapter 3 used n to index the observations to indicate that the data are not ordered,

$$y_n = \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_k x_{nk} + \varepsilon_n. \quad (5.39)$$

Here the observation index is replaced with t to indicate that ordered time-series data are used in the regression,

$$y_t = \beta_1 x_{t1} + \beta_2 x_{t2} + \dots + \beta_k x_{tk} + \varepsilon_t. \quad (5.40)$$

Five assumptions are used to establish the asymptotic distribution of the parameter estimated. Here these assumptions are restated using time-series indices.

Assumption 5.1 (Linearity). *The model specification is linear in \mathbf{x}_t , $y_t = \mathbf{x}_t \beta + \varepsilon_t$.*

Assumption 5.2 (Stationary Ergodicity). *$\{(\mathbf{x}_t, \varepsilon_t)\}$ is a strictly stationary and ergodic sequence.*

Assumption 5.3 (Rank). $E[\mathbf{x}_t' \mathbf{x}_t] = \Sigma_{\mathbf{XX}}$ is non-singular and finite.

Assumption 5.4 (Martingale Difference). *$\{\mathbf{x}_t' \varepsilon_t, \mathcal{F}_{t-1}\}$ is a martingale difference sequence, $E[(x_{j,t} \varepsilon_t)^2] < \infty$ $j = 1, 2, \dots, k$, $t = 1, 2, \dots$ and $\mathbf{S} = V[T^{-1/2} \mathbf{X}' \varepsilon]$ is finite and non singular.*

Assumption 5.5 (Moment Existence). $E[x_{j,t}^4] < \infty$, $j = 1, 2, \dots, k$, $t = 1, 2, \dots$ and $E[\varepsilon_t^2] = \sigma^2 < \infty$, $t = 1, 2, \dots$

Assumption 3.9 may be violated when estimating cross-sectional models using time series data. When this assumption is violated, the scores from the linear regression, $\mathbf{x}_t' \varepsilon_t$ are not martingale difference with respect to the time $t - 1$ information set, \mathcal{F}_{t-1} . The autocorrelation in the scores occurs when the errors from the model, ε_t , have a persistent component that is not explained by the regressors. The MDS assumption featured prominently in two theorems: the asymptotic distribution of $\hat{\beta}$ and the estimation of the covariance of the parameters.

Theorem 5.5. *Under assumptions 3.1 and 3.7 - 3.9*

$$\sqrt{T}(\hat{\beta}_T - \beta) \xrightarrow{d} N(0, \Sigma_{\mathbf{XX}}^{-1} \mathbf{S} \Sigma_{\mathbf{XX}}^{-1}) \quad (5.41)$$

where $\Sigma_{\mathbf{XX}} = E[\mathbf{x}_t' \mathbf{x}_t]$ and $\mathbf{S} = V[T^{-1/2} \mathbf{X}' \varepsilon]$

Theorem 5.6. *Under assumptions 3.1 and 3.7 - 3.10,*

$$\begin{aligned} \hat{\Sigma}_{\mathbf{XX}} &= T^{-1} \mathbf{X}' \mathbf{X} \xrightarrow{p} \Sigma_{\mathbf{XX}} \\ \hat{\mathbf{S}} &= T^{-1} \sum_{n=1}^T e_n^2 \mathbf{x}_n' \mathbf{x}_n \xrightarrow{p} \mathbf{S} \\ &= T^{-1} (\mathbf{X}' \hat{\mathbf{E}} \mathbf{X}) \end{aligned}$$

and

$$\hat{\Sigma}_{\mathbf{XX}}^{-1} \hat{\mathbf{S}} \hat{\Sigma}_{\mathbf{XX}}^{-1} \xrightarrow{p} \Sigma_{\mathbf{XX}}^{-1} \mathbf{S} \Sigma_{\mathbf{XX}}^{-1}$$

where $\hat{\mathbf{E}} = \text{diag}(\hat{\varepsilon}_1^2, \dots, \hat{\varepsilon}_T^2)$ is a matrix with the squared estimated residuals along the diagonal.

When the MDS assumption does not hold, the asymptotic covariance takes a different form that reflects the persistence in the data, and so an alternative estimator is required to estimate the covariance of $\hat{\beta}$. The new estimator is an extended version of White's covariance estimator that accounts for the predictability of the scores ($\mathbf{x}_t' \boldsymbol{\varepsilon}_t$). The correlation in the scores alters the amount of “unique” information available to estimate the parameters. The standard covariance estimator assumes that the scores are uncorrelated with their past and so each contributes its full share to the precision to $\hat{\beta}$. When the scores are autocorrelated, only the unpredictable component of the score is informative about the value of the regression coefficient, and the covariance estimator must account for this change in the available information. Heteroskedasticity Autocorrelation Consistent (HAC) covariance estimators are consistent even in the presence of score autocorrelation.

5.9.1 Estimating the mean with time-series errors

To understand why a HAC estimator is needed, consider estimating the mean in two different setups. In the first, the shock, $\{\boldsymbol{\varepsilon}_t\}$, is assumed to be a white noise process with variance σ^2 . In the second, the shock follows an MA(1) process.

5.9.1.1 White Noise Errors

Suppose the data generating process for Y_t is,

$$Y_t = \mu + \boldsymbol{\varepsilon}_t$$

where $\{\boldsymbol{\varepsilon}_t\}$ is a white noise process. It is simple to show that

$$\mathbb{E}[Y_t] = \mu \text{ and } \mathbb{V}[Y_t] = \sigma^2$$

since the error is a white noise process. Define the sample mean estimator in the usual way,

$$\hat{\mu} = T^{-1} \sum_{t=1}^T Y_t$$

The sample mean is unbiased,

$$\begin{aligned} \mathbb{E}[\hat{\mu}] &= \mathbb{E}\left[T^{-1} \sum_{t=1}^T Y_t\right] \\ &= T^{-1} \sum_{t=1}^T \mathbb{E}[Y_t] \\ &= T^{-1} \sum_{t=1}^T \mu \\ &= \mu. \end{aligned}$$

The variance of the mean estimator exploits the white noise property which ensures $\mathbb{E}[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_j] = 0$ whenever $i \neq j$.

$$\begin{aligned}
V[\hat{\mu}] &= E[(T^{-1} \sum_{t=1}^T Y_t - \mu)^2] \\
&= E[(T^{-1} \sum_{t=1}^T \varepsilon_t)^2] \\
&= E[T^{-2} (\sum_{t=1}^T \varepsilon_t^2 + \sum_{r=1}^T \sum_{s=1, r \neq s}^T \varepsilon_r \varepsilon_s)] \\
&= T^{-2} \sum_{t=1}^T E[\varepsilon_t^2] + T^{-2} \sum_{r=1}^T \sum_{s=1, r \neq s}^T E[\varepsilon_r \varepsilon_s] \\
&= T^{-2} \sum_{t=1}^T \sigma^2 + T^{-2} \sum_{r=1}^T \sum_{s=1, r \neq s}^T 0 \\
&= T^{-2} T \sigma^2 \\
&= \sigma^2 / T,
\end{aligned}$$

and so, $V[\hat{\mu}] = \sigma^2 / T$, the standard result.

5.9.1.2 MA(1) errors

Suppose the model is altered so that the error process ($\{\eta_t\}$) is a mean zero MA(1) constructed from white noise shocks ($\{\varepsilon_t\}$),

$$\eta_t = \theta \varepsilon_{t-1} + \varepsilon_t.$$

The properties of the error are easily derived using the results in Chapter 4. The mean is 0,

$$E[\eta_t] = E[\theta \varepsilon_{t-1} + \varepsilon_t] = \theta E[\varepsilon_{t-1}] + E[\varepsilon_t] = \theta 0 + 0 = 0,$$

and the variance depends on the MA parameter,

$$\begin{aligned}
V[\eta_t] &= E[(\theta \varepsilon_{t-1} + \varepsilon_t)^2] \\
&= E[\theta^2 \varepsilon_{t-1}^2 + 2\theta \varepsilon_{t-1} \varepsilon_t + \varepsilon_t^2] \\
&= E[\theta^2 \varepsilon_{t-1}^2] + 2E[\varepsilon_{t-1} \varepsilon_t] + E[\varepsilon_t^2] \\
&= \theta^2 \sigma^2 + 2 \cdot 0 + \sigma^2 \\
&= \sigma^2(1 + \theta^2).
\end{aligned}$$

The DGP for Y_t is

$$Y_t = \mu + \eta_t,$$

and so the mean and variance of Y_t are

$$E[Y_t] = \mu \text{ and } V[Y_t] = V[\eta_t] = \sigma^2(1 + \theta^2).$$

The sample mean estimator remains unbiased,

$$\hat{\mu} = T^{-1} \sum_{t=1}^T Y_t$$

$$\begin{aligned} E[\hat{\mu}] &= E\left[T^{-1} \sum_{t=1}^T Y_t\right] \\ &= T^{-1} \sum_{t=1}^T E[Y_t] \\ &= T^{-1} \sum_{t=1}^T \mu \\ &= \mu. \end{aligned}$$

The variance of the mean estimator, however, is different, since η_t is autocorrelated, and so $E[\eta_t \eta_{t-1}] \neq 0$.

$$\begin{aligned} V[\hat{\mu}] &= E\left[\left(T^{-1} \sum_{t=1}^T Y_t - \mu\right)^2\right] \\ &= E\left[\left(T^{-1} \sum_{t=1}^T \eta_t\right)^2\right] \\ &= E\left[T^{-2} \left(\sum_{t=1}^T \eta_t^2 + 2 \sum_{t=1}^{T-1} \eta_t \eta_{t+1} + 2 \sum_{t=1}^{T-2} \eta_t \eta_{t+2} + \dots + 2 \sum_{t=1}^2 \eta_t \eta_{t+T-2} + 2 \sum_{t=1}^1 \eta_t \eta_{t+T-1}\right)\right] \\ &= T^{-2} \sum_{t=1}^T E[\eta_t^2] + 2T^{-2} \sum_{t=1}^{T-1} E[\eta_t \eta_{t+1}] + 2T^{-2} \sum_{t=1}^{T-2} E[\eta_t \eta_{t+2}] + \dots + \\ &\quad 2T^{-2} \sum_{t=1}^2 E[\eta_t \eta_{t+T-2}] + 2T^{-2} \sum_{t=1}^1 E[\eta_t \eta_{t+T-1}] \\ &= T^{-2} \sum_{t=1}^T \gamma_0 + 2T^{-2} \sum_{t=1}^{T-1} \gamma_1 + 2T^{-2} \sum_{t=1}^{T-2} \gamma_2 + \dots + 2T^{-2} \sum_{t=1}^2 \gamma_{T-2} + 2T^{-2} \sum_{t=1}^1 \gamma_{T-1} \end{aligned}$$

where $\gamma_0 = E[\eta_t^2] = V[\eta_t]$ and $\gamma_s = E[\eta_t \eta_{t-s}]$. Only γ_0 and γ_1 are non-zero when the error follows an MA(1) process. $\gamma_0 = V[\eta_t] = \sigma^2 (1 + \theta^2)$ and

$$\begin{aligned}
\gamma_1 &= E[\eta_t \eta_{t-1}] \\
&= E[(\theta \varepsilon_{t-1} + \varepsilon_t)(\theta \varepsilon_{t-2} + \varepsilon_{t-1})] \\
&= E[\theta^2 \varepsilon_{t-1} \varepsilon_{t-2} + \theta \varepsilon_{t-1}^2 + \theta \varepsilon_t \varepsilon_{t-2} + \varepsilon_t \varepsilon_{t-1}] \\
&= \theta^2 E[\varepsilon_{t-1} \varepsilon_{t-2}] + \theta E[\varepsilon_{t-1}^2] + \theta E[\varepsilon_t \varepsilon_{t-2}] + E[\varepsilon_t \varepsilon_{t-1}] \\
&= \theta^2 0 + \theta \sigma^2 + \theta 0 + 0 \\
&= \theta \sigma^2.
\end{aligned}$$

The remaining autocovariance are all 0 since $\gamma_s = 0$, $s > Q$ in a MA(Q). Returning to the variance of $\hat{\mu}$,

$$\begin{aligned}
V[\hat{\mu}] &= T^{-2} \sum_{t=1}^T \gamma_0 + 2T^{-2} \sum_{t=1}^{T-1} \gamma_1 \\
&= T^{-2} T \gamma_0 + 2T^{-2} (T-1) \gamma_1 \\
&\approx \frac{\gamma_0 + 2\gamma_1}{T}.
\end{aligned} \tag{5.42}$$

When the errors are autocorrelated, the usual mean estimator has a different variance that reflects the dependence in the errors. Importantly, the usual estimator variance is no longer correct and $V[\hat{\mu}] \neq \gamma_0/T$.

This simple illustration captures the key idea that underlies the Newey-West variance estimator,

$$\hat{\sigma}_{NW}^2 = \hat{\gamma}_0 + 2 \sum_{l=1}^L \left(1 - \frac{l}{L+1}\right) \hat{\gamma}_l.$$

When $L=1$, the only weight is $1 - 1/2 = 1/2$ and $\hat{\sigma}_{NW}^2 = \hat{\gamma}_0 + \hat{\gamma}_1$, which is different from the variance in the MA(1) error example. However as L increases, the weight on γ_1 converges to 1 since $\lim_{L \rightarrow \infty} 1 - \frac{1}{L+1} = 1$. The Newey-West variance estimator asymptotically includes all of the autocovariance in the variance, $\gamma_0 + 2\gamma_1$, and when L grows large,

$$\hat{\sigma}_{NW}^2 \rightarrow \gamma_0 + 2\gamma_1.$$

The variance of the estimated mean can be consistently estimated using $\hat{\sigma}_{NW}^2$ as

$$V[\hat{\mu}] = \frac{\gamma_0 + 2\gamma_1}{T} \approx \frac{\hat{\sigma}_{NW}^2}{T}.$$

As a general principle, the variance of the sum is the sum of the variances only true when the errors are uncorrelated. HAC covariance estimators account for time-series dependence and lead to correct inference as long as L grows with the sample size.²¹

It is tempting to estimate eq. (5.42) using the natural estimator $\hat{\sigma}_{HAC}^2 = \hat{\gamma}_0 + 2\hat{\gamma}_1/T$. This estimator is not guaranteed to be positive in finite samples, an in general unweighted estimators of the form

²¹Allowing L to grow at the rate $T^{1/3}$ is optimal in a certain sense related to testing.

$\hat{\sigma}_{HAC}^2 = \hat{\gamma}_0 + 2\hat{\gamma}_1 + 2\hat{\gamma}_2 + \dots + 2\hat{\gamma}_L$, may be negative. The Newey-West variance estimator, $\hat{\sigma}_{NW}^2$, is guaranteed to be positive for any L . The weights that scale the autocovariances, $(1 - \frac{l}{L+1})$, alter the estimator and ensure that the estimate is positive.

5.9.2 Estimating the variance of $\hat{\beta}$ when the errors are autocorrelated

There are two solutions to modeling cross-sectional data that have autocorrelated errors. The direct method is to alter the cross-sectional model to capture the time-series variation by including both contemporaneous effects of \mathbf{x}_t as well as lagged values of y_t (and possibly lags of \mathbf{x}_t). This approach needs to include sufficient lags so that the errors are white noise. If the dependence is fully modeled, then White's heteroskedasticity (but not autocorrelation) consistent covariance estimator is consistent, and there is no need for a more complex covariance estimator.

The second approach modifies the covariance estimator to account for the dependence in the data. The key insight in White's estimator of \mathbf{S} ,

$$\hat{\mathbf{S}} = T^{-1} \sum_{t=1}^T e_t^2 \mathbf{x}'_t \mathbf{x}_t,$$

is that this form explicitly captures the dependence between the e_t^2 and $\mathbf{x}'_t \mathbf{x}_t$. Heteroskedasticity Autocorrelation Consistent estimators work similarly by capturing both the dependence between the e_t^2 and $\mathbf{x}'_t \mathbf{x}_t$ (heteroskedasticity) and the dependence between the $\mathbf{x}_t e_t$ and $\mathbf{x}_{t-j} e_{t-j}$ (autocorrelation). HAC estimators of the score covariance in linear regressions use the same structure, and

$$\begin{aligned} \hat{\mathbf{S}}_{HAC} &= T^{-1} \left(\sum_{t=1}^T e_t^2 \mathbf{x}'_t \mathbf{x}_t + \sum_{l=1}^L w_l \left(\sum_{s=l+1}^T e_s e_{s-l} \mathbf{x}'_s \mathbf{x}_{s-l} + \sum_{q=l+1}^T e_{q-l} e_q \mathbf{x}'_{q-l} \mathbf{x}_q \right) \right) \\ &= \hat{\Gamma}_0 + \sum_{l=1}^L w_l (\hat{\Gamma}_l + \hat{\Gamma}_{-l}) \\ &= \hat{\Gamma}_0 + \sum_{l=1}^L w_l (\hat{\Gamma}_l + \hat{\Gamma}'_l) \end{aligned} \quad (5.43)$$

where $\{w_l\}$ are a set of weights. The Newey-West estimator uses $w_l = 1 - \frac{l}{L+1}$ and is always positive semi-definite. Other estimators alter the weights and have different finite-sample properties.

5.A Cointegration in a trivariate VAR

This section details how to:

- determine whether a trivariate VAR is cointegrated;
- determine the number of cointegrating vectors in a cointegrated system; and
- decompose the π matrix into α , the adjustment coefficient, and β , the cointegrating vectors.

5.A.1 Stationary VAR

Consider the VAR(1),

$$\begin{bmatrix} x_t \\ y_t \\ z_t \end{bmatrix} = \begin{bmatrix} .9 & -.4 & .2 \\ .2 & .8 & -.3 \\ .5 & .2 & .1 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \end{bmatrix}.$$

The eigenvalues of the parameter matrix determine the stationarity properties of this VAR process. If the eigenvalues are all less than one in modulus, then the VAR(1) is stationary. This is the case here, and the eigenvalues are 0.97, 0.62, and 0.2. An alternative method is to transform the model into an VECM

$$\begin{aligned} \begin{bmatrix} \Delta x_t \\ \Delta y_t \\ \Delta z_t \end{bmatrix} &= \left(\begin{bmatrix} .9 & -.4 & .2 \\ .2 & .8 & -.3 \\ .5 & .2 & .1 \end{bmatrix} - \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right) \begin{bmatrix} x_{t-1} \\ y_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \end{bmatrix} \\ \begin{bmatrix} \Delta x_t \\ \Delta y_t \\ \Delta z_t \end{bmatrix} &= \begin{bmatrix} -.1 & -.4 & .2 \\ .2 & -.2 & -.3 \\ .5 & .2 & -.9 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \end{bmatrix} \\ \Delta \mathbf{w}_t &= \boldsymbol{\pi} \mathbf{w}_t + \varepsilon_t \end{aligned}$$

where \mathbf{w}_t is a vector composed of x_t , y_t and z_t . The rank of the parameter matrix $\boldsymbol{\pi}$ can be determined by transforming it into row-echelon form.

$$\begin{bmatrix} -0.1 & -0.4 & 0.2 \\ 0.2 & -0.2 & -0.3 \\ 0.5 & 0.2 & -0.9 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 4 & -2 \\ 0.2 & -0.2 & -0.3 \\ 0.5 & 0.2 & -0.9 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 4 & -2 \\ 0 & -1 & 0.1 \\ 0 & -1.8 & 0.1 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 4 & -2 \\ 0 & 1 & -0.1 \\ 0 & -1.8 & 0.1 \end{bmatrix} \\ \Rightarrow \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -0.1 \\ 0 & 0 & -0.08 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -0.1 \\ 0 & 0 & 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Since the $\boldsymbol{\pi}$ matrix is full rank, the system must be stationary. This method is equivalent to computing the eigenvalues of the parameter matrix in the VAR.

5.A.2 Independent Unit Roots

Consider the simple VAR

$$\begin{bmatrix} x_t \\ y_t \\ z_t \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \end{bmatrix}.$$

The eigenvalues of the coefficient matrix are all 1 and the VECM is

$$\begin{aligned} \begin{bmatrix} \Delta x_t \\ \Delta y_t \\ \Delta z_t \end{bmatrix} &= \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right) \begin{bmatrix} x_{t-1} \\ y_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \end{bmatrix} \\ \begin{bmatrix} \Delta x_t \\ \Delta y_t \\ \Delta z_t \end{bmatrix} &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \end{bmatrix} \end{aligned}$$

and the rank of π is clearly 0, so this system contains three independent unit roots. This structure also applies to higher order nonstationary VAR models that contain independent unit root processes – the coefficient matrix in the VECM is always rank 0 when the system contains as many distinct unit roots as variables.

5.A.3 Cointegrated with one cointegrating relationship

Consider the VAR(1),

$$\begin{bmatrix} x_t \\ y_t \\ z_t \end{bmatrix} = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ -0.16 & 1.08 & 0.08 \\ 0.36 & -0.18 & 0.82 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \end{bmatrix}.$$

The eigenvalues of the parameter matrix are 1, 1 and .7. The VECM form of this model is

$$\begin{bmatrix} \Delta x_t \\ \Delta y_t \\ \Delta z_t \end{bmatrix} = \left(\begin{bmatrix} 0.8 & 0.1 & 0.1 \\ -0.16 & 1.08 & 0.08 \\ 0.36 & -0.18 & 0.82 \end{bmatrix} - \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right) \begin{bmatrix} x_{t-1} \\ y_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \end{bmatrix}$$

$$\begin{bmatrix} \Delta x_t \\ \Delta y_t \\ \Delta z_t \end{bmatrix} = \begin{bmatrix} -0.2 & 0.1 & 0.1 \\ -0.16 & 0.08 & 0.08 \\ 0.36 & -0.18 & -0.18 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \end{bmatrix}.$$

The eigenvalues of π are 0, 0 and $-.3$, and so $\text{rank}(\pi) = 1$. Recall that the number of cointegrating vectors is the rank of π in a cointegrated system. In this example, there is one cointegrating vector, which can be computed by transforming π into row-echelon form,

$$\begin{bmatrix} -0.2 & 0.1 & 0.1 \\ -0.16 & 0.08 & 0.08 \\ 0.36 & -0.18 & -0.18 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & -0.5 & -0.5 \\ -0.16 & 0.08 & 0.08 \\ 0.36 & -0.18 & -0.18 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & -0.5 & -0.5 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

The cointegrating vector is $\beta = [1 \ -0.5 \ -0.5]'$ and α is found by noting that

$$\pi = \alpha\beta' = \begin{bmatrix} \alpha_1 & -\frac{1}{2}\alpha_1 & -\frac{1}{2}\alpha_1 \\ \alpha_2 & -\frac{1}{2}\alpha_2 & -\frac{1}{2}\alpha_2 \\ \alpha_3 & -\frac{1}{2}\alpha_3 & -\frac{1}{2}\alpha_3 \end{bmatrix},$$

so that $\alpha = [-.2 \ -.16 \ .36]'$ is the first column of π .

5.A.4 Cointegrated with two cointegrating relationships

Consider the VAR(1),

$$\begin{bmatrix} x_t \\ y_t \\ z_t \end{bmatrix} = \begin{bmatrix} 0.3 & 0.4 & 0.3 \\ 0.1 & 0.5 & 0.4 \\ 0.2 & 0.2 & 0.6 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \end{bmatrix}.$$

The eigenvalues of the parameter matrix are 1, $.2+.1i$ and $.2-.1i$, which have complex moduli of 1, .223 and .223, respectively. The VECM form of this model is

$$\begin{bmatrix} \Delta x_t \\ \Delta y_t \\ \Delta z_t \end{bmatrix} = \left(\begin{bmatrix} 0.3 & 0.4 & 0.3 \\ 0.1 & 0.5 & 0.4 \\ 0.2 & 0.2 & 0.6 \end{bmatrix} - \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right) \begin{bmatrix} x_{t-1} \\ y_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \end{bmatrix}$$

$$\begin{bmatrix} \Delta x_t \\ \Delta y_t \\ \Delta z_t \end{bmatrix} = \begin{bmatrix} -0.7 & 0.4 & 0.3 \\ 0.1 & -0.5 & 0.4 \\ 0.2 & 0.2 & -0.4 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \\ z_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \\ \varepsilon_{3,t} \end{bmatrix}.$$

The eigenvalues of π are 0, $-0.8+0.1i$ and $-0.8-0.1i$, and so $\text{rank}(\pi) = 2$. The number of cointegrating vectors is the rank of π . One set of cointegrating vectors can be found by transforming π into row-echelon form²²,

$$\begin{bmatrix} -0.7 & 0.4 & 0.3 \\ 0.1 & -0.5 & 0.4 \\ 0.2 & 0.2 & -0.4 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & -0.57143 & -0.42857 \\ 0.1 & -0.5 & 0.4 \\ 0.2 & 0.2 & -0.4 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & -0.57143 & -0.42857 \\ 0 & -0.44286 & 0.44286 \\ 0 & 0.31429 & -0.31429 \end{bmatrix} \Rightarrow$$

$$\begin{bmatrix} 1 & -0.57143 & -0.42857 \\ 0 & 1 & -1 \\ 0 & 0.31429 & -0.31429 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix}$$

β is the transpose of first two rows of the row-echelon form,

$$\beta = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{bmatrix}$$

α is found using the relationship

$$\pi = \alpha\beta' = \begin{bmatrix} \alpha_{11} & \alpha_{12} & -\alpha_{11} - \alpha_{12} \\ \alpha_{21} & \alpha_{22} & -\alpha_{21} - \alpha_{22} \\ \alpha_{31} & \alpha_{32} & -\alpha_{31} - \alpha_{32} \end{bmatrix},$$

and so α is the first two columns of π ,

$$\alpha = \begin{bmatrix} -0.7 & 0.4 \\ 0.1 & -0.5 \\ 0.2 & 0.2 \end{bmatrix}.$$

²²The cointegrating vectors are only defined up to an arbitrary normalization. Any set of cointegrating vectors β and be used to create a different set by multiplying by a k by k full-rank matrix \mathbf{A} so that $\tilde{\beta} = \mathbf{A}\beta$ is also a cointegrating vector.

Exercises

Shorter Questions

Problem 5.1. Under what conditions are two random variables cointegrated?

Problem 5.2. Suppose $\mathbf{y}_t = \Phi_0 + \Phi_1 \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t$ where \mathbf{y}_t is a k by 1 vector values variable and Φ_0 and Φ_1 are conformable. What are the 1 and 2 step forecasts from this model?

Longer Questions

Exercise 5.1. Consider the estimation of a mean where the errors are a white noise process.

- Show that the usual mean estimator is unbiased and derive its variance *without assuming the errors are i.i.d.*

Now suppose error process follows an MA(1) so that $\boldsymbol{\varepsilon}_t = \mathbf{v}_t + \theta_1 \mathbf{v}_{t-1}$ where \mathbf{v}_t is a WN process.

- Show that the usual mean estimator is still unbiased and derive the variance of the mean.

Suppose that $\{\eta_{1,t}\}$ and $\{\eta_{2,t}\}$ are two sequences of uncorrelated i.i.d. standard normal random variables.

$$\begin{aligned} x_t &= \eta_{1,t} + \theta_{11} \eta_{1,t-1} + \theta_{12} \eta_{2,t-1} \\ y_t &= \eta_{2,t} + \theta_{21} \eta_{1,t-1} + \theta_{22} \eta_{2,t-1} \end{aligned}$$

- What are $E_t[x_{t+1}]$ and $E_t[x_{t+2}]$?
- Define the autocovariance matrix of a vector process.
- Compute the autocovariance matrix Γ_j for $j = 0, \pm 1$.

Exercise 5.2. Consider an AR(1)

- What are the two types of stationarity? Provide precise definitions.
- Which of the following bivariate Vector Autoregressions are stationary? If they are not stationary are they cointegrated, independent unit roots or explosive? Assume

$$\begin{bmatrix} \boldsymbol{\varepsilon}_{1t} \\ \boldsymbol{\varepsilon}_{2t} \end{bmatrix} \stackrel{\text{i.i.d.}}{\sim} N(\mathbf{0}, \mathbf{I}_2)$$

Recall that the eigenvalues values of a 2×2 non-triangular matrix

$$\boldsymbol{\pi} = \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix}$$

can be solved using the two-equation, two-unknowns system $\lambda_1 + \lambda_2 = \pi_{11} + \pi_{22}$ and $\lambda_1 \lambda_2 = \pi_{11} \pi_{22} - \pi_{12} \pi_{21}$.

(a)

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} 1.4 & .4 \\ -.6 & .4 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}$$

(b)

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}$$

(c)

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} .8 & 0 \\ .2 & .4 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}$$

- iii. What are spurious regression and balance?
- iv. Why is spurious regression a problem?
- v. Briefly outline the steps needed to test for a spurious regression in

$$y_t = \beta_1 + \beta_2 x_t + \varepsilon_t.$$

Exercise 5.3. Consider the AR(2)

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t.$$

- i. Rewrite the model with Δy_t on the left-hand side and y_{t-1} and Δy_{t-1} on the right-hand side.
- ii. What restrictions are needed on ϕ_1 and ϕ_2 for this model to collapse to an AR(1) in the first differences?
- iii. When the model collapses, what does this imply about y_t ?

Consider the VAR(1)

$$\begin{aligned} x_t &= x_{t-1} + \varepsilon_{1,t} \\ y_t &= \beta x_{t-1} + \varepsilon_{2,t} \end{aligned}$$

where $\{\varepsilon_t\}$ is a vector white noise process.

- i. Are x_t and y_t cointegrated?
- ii. Write this model in error correction form.

Consider the VAR(1)

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} 0.4 & 0.3 \\ 0.8 & 0.6 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}$$

where $\{\varepsilon_t\}$ is a vector white noise process.

- i. Verify that x_t and y_t are cointegrated.

- ii. Write this model in error correction form.
- iii. Compute the speed of adjustment coefficient α and the cointegrating vector β where the β on x_t is normalized to 1.

Exercise 5.4. Data on interest rates on US government debt is collected for 3-month (*3MO*) T-bills, and 3-year (*3YR*) and 10-year (*10YR*) bonds from 1957 until 2009. Three transformed variables are defined using these three series:

Level	<i>3MO</i>
Slope	<i>10YR - 3MO</i>
Curvature	$(10YR - 3YR) - (3YR - 3MO)$

- i. In terms of VAR analysis, does it matter whether the original data or the level-slope-curvature model is fit? Hint: Think about reparameterizations between the two.

Granger Causality analysis is performed on this set, and the p-values are

	Level _{t-1}	Slope _{t-1}	Curvature _{t-1}
Level _t	0.000	0.244	0.000
Slope _t	0.000	0.000	0.000
Curvature _t	0.000	0.000	0.000
All (excl. self)	0.000	0.000	0.000

- ii. Interpret this table.
- iii. When constructing impulse response graphs the selection of the covariance of the shocks is important. Outline the alternatives and describe situations when each may be preferable.
- iv. Figure 5.7 contains the impulse response curves for this model. Interpret the graph. Also, comment on why the impulse responses can all be significantly different from 0 in light of the Granger Causality table.
- v. Why are some of the lag-0 impulses precisely 0.0?

Exercise 5.5. Answer the following questions:

- i. Consider the AR(2)

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$$

- (a) Rewrite the model with Δy_t on the left-hand side and y_{t-1} and Δy_{t-1} on the right-hand side.
- (b) What restrictions are needed on ϕ_1 and ϕ_2 for this model to collapse to an AR(1) in the first differences?
- (c) When the model collapses, what does this imply about y_t ?

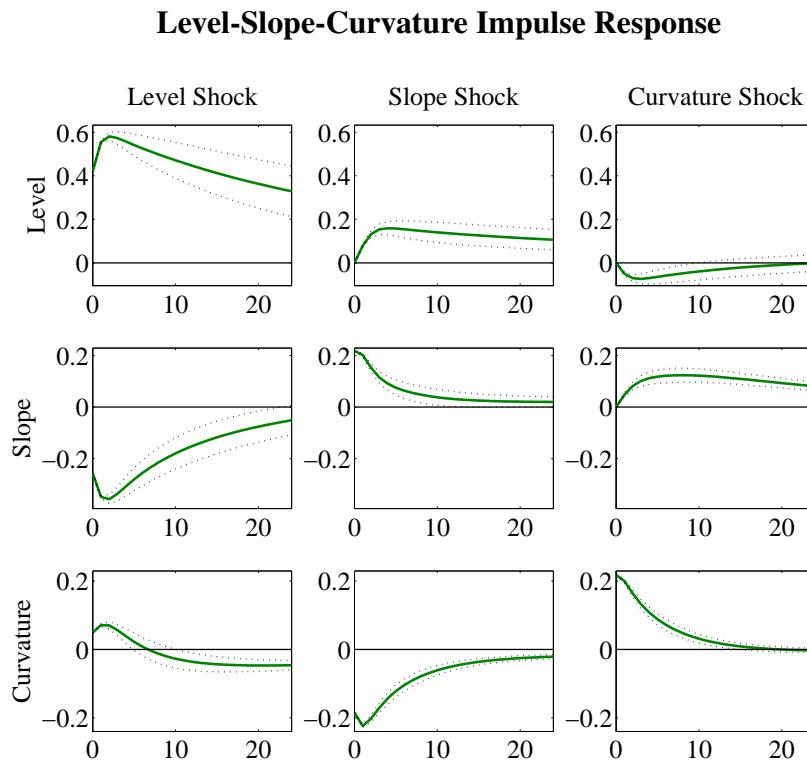


Figure 5.7: Impulse response functions and 95% confidence intervals for the level-slope-curvature exercise.

ii. Consider the VAR(1)

$$\begin{aligned} x_t &= x_{t-1} + \varepsilon_{1,t} \\ y_t &= \beta x_{t-1} + \varepsilon_{2,t} \end{aligned}$$

where $\{\varepsilon_t\}$ is a vector white noise process.

- (a) Are x_t and y_t cointegrated?
- (b) Write this model in error correction form.

iii. Consider the VAR(1)

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} 0.625 & -0.3125 \\ -0.75 & 0.375 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}$$

where $\{\varepsilon_t\}$ is a vector white noise process.

- (a) Verify that x_t and y_t are cointegrated.
- (b) Write this model in error correction form.

- (c) Compute the speed of adjustment coefficient α and the cointegrating vector β where the β on x_t is normalized to 1.

Exercise 5.6. Consider the estimation of a mean where the errors are a white noise process.

- i. Show that the usual mean estimator is unbiased and derive its variance *without assuming the errors are i.i.d.*

Now suppose error process follows an AR(1) so that $y_t = \mu + \varepsilon_t$ and $\varepsilon_t = \rho \varepsilon_{t-1} + v_t$ where $\{v_t\}$ is a WN process.

- ii. Show that the usual mean estimator is still unbiased and derive the variance of the sample mean.
 iii. What is Granger Causality and how is it useful in Vector Autoregression analysis? Be as specific as possible.

Suppose that $\{\eta_{1,t}\}$ and $\{\eta_{2,t}\}$ are two sequences of uncorrelated i.i.d. standard normal random variables.

$$\begin{aligned}x_t &= \eta_{1,t} + \theta_{11}\eta_{1,t-1} + \theta_{12}\eta_{2,t-1} \\y_t &= \eta_{2,t} + \theta_{21}\eta_{1,t-1} + \theta_{22}\eta_{2,t-1}\end{aligned}$$

- iv. Define the autocovariance matrix of a vector process.
 v. Compute the autocovariance matrix Γ_j for $j = 0, \pm 1$.
 vi. The AIC, HQIC, and BIC are computed for a bivariate VAR with lag length ranging from 0 to 12 and are in the table below. Which model is selected by each criterion?

Lag Length	AIC	HQIC	BIC
0	2.1916	2.1968	2.2057
1	0.9495	0.9805	1.0339
2	0.9486	1.0054	1.1032
3	0.9716	1.0542	1.1965
4	0.9950	1.1033	1.2900
5	1.0192	1.1532	1.3843
6	1.0417	1.2015	1.4768
7	1.0671	1.2526	1.5722
8	1.0898	1.3010	1.6649
9	1.1115	1.3483	1.7564
10	1.1331	1.3956	1.8478
11	1.1562	1.4442	1.9406
12	1.1790	1.4926	2.0331

Exercise 5.7. Consider the VAR(1)

$$\begin{aligned}x_t &= x_{t-1} + \varepsilon_{1,t} \\y_t &= \beta x_{t-1} + \varepsilon_{2,t}\end{aligned}$$

where $\{\varepsilon_t\}$ is a vector white noise process.

- i. Are x_t and y_t cointegrated?
- ii. Write this model in error correction form.

Exercise 5.8. Answer the following questions.

- i. Describe two methods for determining the number of lags to use in a VAR(P)
- ii. Consider the VAR(P)

$$\mathbf{y}_t = \Phi_1 \mathbf{y}_{t-1} + \Phi_2 \mathbf{y}_{t-2} + \varepsilon_t.$$

Write this VAR in companion form. Under what conditions is this process stationary?

- iii. For the remainder of the question, consider the 2-dimentional VAR(1)

$$\mathbf{y}_t = \Phi_1 \mathbf{y}_{t-1} + \varepsilon_t.$$

Define Granger Causality and explain what conditions on Φ_1 are needed for no series in \mathbf{y}_t to Granger cause any other series in \mathbf{y}_t .

- iv. Define cointegration in this system.
- v. What conditions on Φ_1 are required for the VAR(1) to be cointegrated?
- vi. Write the VAR(1) in error correction form.
- vii. In this setup, describe how to test for cointegration using the Engle-Granger method.

Exercise 5.9. Consider a VAR(1)

$$\mathbf{y}_t = \Phi_1 \mathbf{y}_{t-1} + \varepsilon_t$$

- i. What are the impulses in this model?
- ii. Define cointegration for this model.
- iii. What conditions on the eigenvalues of Φ_1 are required for cointegration to be present?
- iv. Consider a 2-dimensional VAR(1) written in error correction form

$$\Delta \mathbf{y}_t = \Pi \mathbf{y}_{t-1} + \varepsilon_t.$$

Assume each of the variables in \mathbf{y}_t are $I(1)$. What conditions on the rank of Π must hold when:

- (a) \mathbf{y}_{t-1} are stationary
- (b) \mathbf{y}_{t-1} are cointegrated

- (c) \mathbf{y}_{t-1} are random walks
- v. Define spurious regression. Why is this a problem?

Exercise 5.10. Consider the VAR(P)

$$\mathbf{y}_t = \Phi_1 \mathbf{y}_{t-1} + \Phi_2 \mathbf{y}_{t-2} + \varepsilon_t$$

- i. Write this in companion form. Under what conditions is the VAR(P) stationary?

- ii. Consider the 2-dimentional VAR

$$\mathbf{y}_t = \Phi_1 \mathbf{y}_{t-1} + \varepsilon_t$$

- (a) What conditions on Φ_1 are required for the VAR(1) to have cointegration?
- (b) Describe how to test for cointegration using the Engle-Granger method.

Exercise 5.11. Consider a VAR(1)

$$\mathbf{y}_t = \Phi_1 \mathbf{y}_{t-1} + \varepsilon_t$$

- i. What is an impulse response function for this model?
- ii. Define cointegration for this model.
- iii. What conditions on the eigenvalues of Φ_1 are required for cointegration to be present?
- iv. Consider a 2-dimensional VAR(1) written in error correction form

$$\Delta \mathbf{y}_t = \Pi \mathbf{y}_{t-1} + \varepsilon_t.$$

Assume each of the variables in \mathbf{y}_t are $I(1)$. What conditions on the rank of Π must hold when:

- (a) \mathbf{y}_{t-1} are stationary
- (b) \mathbf{y}_{t-1} are cointegrated
- (c) \mathbf{y}_{t-1} are random walks
- v. Define spurious regression. Why is this a problem?

Exercise 5.12. Answer the following questions.

- i. Describe two methods for determining the number of lags to use in a VAR(P)
- ii. Consider the VAR(P)

$$\mathbf{y}_t = \Phi_1 \mathbf{y}_{t-1} + \Phi_2 \mathbf{y}_{t-2} + \varepsilon_t.$$

Write this in companion form. Under what conditions is the VAR(P) stationary?

- iii. For the remainder of the question, consider the 2-dimentional VAR(1)

$$\mathbf{y}_t = \Phi_1 \mathbf{y}_{t-1} + \varepsilon_t.$$

Define Granger Causality and explain what conditions on Φ_1 are needed for no series in \mathbf{y}_t to Granger cause any other series in \mathbf{y}_t .

- iv. Define cointegration in this system.
- v. What conditions on Φ_1 are required for the VAR(1) to have cointegration?
- vi. Write the VAR(1) in error correction form.
- vii. In this setup, describe how to test for cointegration using the Engle-Granger method.

Chapter 6

Generalized Method Of Moments (GMM)

Note: The primary reference text for these notes is Hall (2005). Alternative, but less comprehensive, treatments can be found in chapter 14 of Hamilton (1994) or some sections of chapter 4 of Greene (2007). For an excellent perspective of GMM from a finance point of view, see chapters 10, 11 and 13 in Cochrane (2001).

Generalized Moethod of Moments is a broadly applicable parameter estimation strategy which nests the classic method of moments, linear regression, maximum likelihood. This chapter discusses the specification of moment conditions – the building blocks of GMM estimations, estimation, inference and specificatrion testing. These ideas are illustrated through three examples: estimation of a consumption asset pricing model, linear factors models and stochastic volatility.

Generalized Method of Moments (GMM) is an estimation procedure that allows economic models to be specified while avoiding often unwanted or unnecessary assumptions, such as specifying a particular distribution for the errors. This lack of structure means GMM is widely applicable, although this generality comes at the cost of a number of issues, the most important of which is questionable small sample performance. This chapter introduces the GMM estimation strategy, discuss specification, estimation, inference and testing.

6.1 Classical Method of Moments

The classical method of moments, or simply method of moments, uses sample moments to estimate unknown parameters. For example, suppose a set of T observations, y_1, \dots, y_T are i.i.d. Poisson with intensity parameter λ . Since $E[y_t] = \lambda$, a natural method to estimate the unknown parameter is to use the sample average,

$$\hat{\lambda} = T^{-1} \sum_{t=1}^T y_t \tag{6.1}$$

which converges to λ as the sample size grows large. In the case of Poisson data , the mean is not the only moment which depends on λ , and so it is possible to use other moments to learn about the intensity. For example the variance $V[y_t] = \lambda$, also depends on λ and so $E[y_t^2] = \lambda^2 + \lambda$. This can be used estimate to lambda since

$$\lambda + \lambda^2 = E \left[T^{-1} \sum_{t=1}^T y_t^2 \right] \quad (6.2)$$

and, using the quadratic formula, an estimate of λ can be constructed as

$$\hat{\lambda} = \frac{-1 + \sqrt{1 + 4\bar{y}^2}}{2} \quad (6.3)$$

where $\bar{y}^2 = T^{-1} \sum_{t=1}^T y_t^2$. Other estimators for λ could similarly be constructed by computing higher order moments of y_t .¹ These estimators are method of moments estimators since they use sample moments to estimate the parameter of interest. Generalized Method of Moments (GMM) extends the classical setup in two important ways. The first is to formally treat the problem of having two or more moment conditions which have information about unknown parameters. GMM allows estimation and inference in systems of Q equations with P unknowns, $P \leq Q$. The second important generalization of GMM is that quantities other than sample moments can be used to estimate the parameters. GMM exploits laws of large numbers and central limit theorems to establish regularity conditions for many different “moment conditions” that may or may not actually be moments. These two changes produce a class of estimators that is broadly applicable. Section 6.7 shows that the classical method of moments, ordinary least squares and maximum likelihood are all special cases of GMM.

6.2 Examples

Three examples will be used throughout this chapter. The first is a simple consumption asset pricing model. The second is the estimation of linear asset pricing models and the final is the estimation of a stochastic volatility model.

6.2.1 Consumption Asset Pricing

GMM was originally designed as a solution to a classic problem in asset pricing: how can a consumption based model be estimated without making strong assumptions on the distribution of returns? This example is based on Hansen and Singleton (1982), a model which builds on Lucas (1978).

The classic consumption based asset pricing model assumes that a representative agent maximizes the conditional expectation of their lifetime discounted utility,

$$E_t \left[\sum_{i=0}^{\infty} \beta^i U(c_{t+i}) \right] \quad (6.4)$$

where β is the discount rate (rate of time preference) and $U(\cdot)$ is a strictly concave utility function. Agents allocate assets between N risky assets and face the budget constraint

¹The quadratic formula has two solutions. It is simple to verify that the other solution, $\frac{-1 - \sqrt{1 + 4\bar{y}^2}}{2}$, is negative and so cannot be the intensity of a Poisson process.

$$c_t + \sum_{j=1}^N p_{j,t} q_{j,t} = \sum_{j=1}^N R_{j,t} q_{j,t-m_j} + w_t \quad (6.5)$$

where c_t is consumption, $p_{j,t}$ and $q_{j,t}$ are price and quantity of asset j , $j = 1, 2, \dots, N$, $R_{j,t}$ is the time t payoff of holding asset j purchased in period $t - m_j$, $q_{j,t-m_j}$ is the amount purchased in period $t - m_j$ and w_t is real labor income. The budget constraint requires that consumption plus asset purchases (LHS) is equal to portfolio wealth plus labor income. Solving this model produces a standard Euler equation,

$$p_{j,t} U'(c_t) = \beta^{m_j} E_t [R_{j,t+m_j} U'(c_{t+m_j})] \quad (6.6)$$

which is true for all assets and all time periods. This Euler equation states that the utility foregone by purchasing an asset at $p_{j,t}$ must equal the discounted expected utility gained from holding that asset in period $t + m_j$. The key insight of Hansen and Singleton (1982) is that this simple condition has many testable implications, mainly that

$$E_t \left[\beta^{m_j} \left(\frac{R_{j,t+m_j}}{p_{j,t}} \right) \left(\frac{U'(c_{t+m_j})}{U'(c_t)} \right) \right] - 1 = 0 \quad (6.7)$$

Note that $\frac{R_{j,t+m_j}}{p_{j,t}}$ is the gross rate of return for asset j (1 plus the net rate of return). Since the Euler equation holds for all time horizons, it is simplest to reduce it to a one-period problem. Defining $r_{j,t+1}$ to be the net rate of return one period ahead for asset j ,

$$E_t \left[\beta (1 + r_{j,t+1}) \left(\frac{U'(c_{t+1})}{U'(c_t)} \right) \right] - 1 = 0 \quad (6.8)$$

which provides a simple testable implication of this model. This condition must be true for any asset j which provides a large number of testable implications by replacing the returns of one series with those of another. Moreover, the initial expectation is conditional which produces further implications for the model. Not only is the Euler equation required to have mean zero, it must be uncorrelated with any time t instrument z_t , and so it must also be the case that

$$E \left[\left(\beta (1 + r_{j,t+1}) \left(\frac{U'(c_{t+1})}{U'(c_t)} \right) - 1 \right) z_t \right] = 0. \quad (6.9)$$

The use of conditioning information can be used to construct a huge number of testable restrictions. This model is completed by specifying the utility function to be CRRA,

$$U(c_t) = \frac{c_t^{1-\gamma}}{1-\gamma} \quad (6.10)$$

$$U'(c_t) = c_t^{-\gamma} \quad (6.11)$$

where γ is the coefficient of relative risk aversion. With this substitution, the testable implications are

$$E \left[\left(\beta (1 + r_{j,t+1}) \left(\frac{c_{t+1}}{c_t} \right)^{-\gamma} - 1 \right) z_t \right] = 0 \quad (6.12)$$

where z_t is any t available instrument (including a constant, which will produce an unconditional restriction).

6.2.2 Linear Factor Models

Linear factor models are widely popular in finance due to their ease of estimation using the Fama and MacBeth (1973) methodology and the Shanken (1992) correction. However, Fama-MacBeth, even with the correction, has a number of problems; the most important is that the assumptions underlying the Shanken correction are not valid for heteroskedastic asset pricing models and so the modified standard errors are not consistent. GMM provides a simple method to estimate linear asset pricing models and to make correct inference under weaker conditions than those needed to derive the Shanken correction. Consider the estimation of the parameters of the CAPM using two assets. This model contains three parameters: the two β s, measuring the risk sensitivity, and λ_m , the market price of risk. These two parameters are estimated using four equations,

$$\begin{aligned} r_{1t}^e &= \beta_1 r_{mt}^e + \varepsilon_{1t} \\ r_{2t}^e &= \beta_2 r_{mt}^e + \varepsilon_{2t} \\ r_{1t}^e &= \beta_1 \lambda^m + \eta_{1t} \\ r_{2t}^e &= \beta_2 \lambda^m + \eta_{2t} \end{aligned} \tag{6.13}$$

where $r_{j,t}^e$ is the excess return to asset j , r_{mt}^e is the excess return to the market and $\varepsilon_{j,t}$ and $\eta_{j,t}$ are errors.

These equations should look familiar; they are the Fama-Macbeth equations. The first two – the “time-series” regressions – are initially estimated using OLS to find the values for β_j , $j = 1, 2$ and the last two – the “cross-section” regression – are estimated conditioning on the first stage β s to estimate the price of risk. The Fama-MacBeth estimation procedure can be used to generate a set of equations that should have expectation zero at the correct parameters. The first two come from the initial regressions (see chapter 3),

$$\begin{aligned} (r_{1t}^e + \beta_1 r_{mt}^e) r_{mt}^e &= 0 \\ (r_{2t}^e + \beta_2 r_{mt}^e) r_{mt}^e &= 0 \end{aligned} \tag{6.14}$$

and the last two come from the second stage regressions

$$\begin{aligned} r_{1t}^e - \beta_1 \lambda^m &= 0 \\ r_{2t}^e - \beta_2 \lambda^m &= 0 \end{aligned} \tag{6.15}$$

This set of equations consists 3 unknowns and four equations and so cannot be directly estimated using least squares. One of the main advantages of GMM is that it allows estimation in systems where the number of unknowns is smaller than the number of moment conditions, and to test whether the moment conditions hold (all conditions not significantly different from 0).

6.2.3 Stochastic Volatility Models

Stochastic volatility is an alternative framework to ARCH for modeling conditional heteroskedasticity. The primary difference between the two is the inclusion of 2 (or more) shocks in stochastic volatility models. The inclusion of the additional shock makes standard likelihood-based methods, like those used to estimate ARCH-family models, infeasible. GMM was one of the first methods used to estimate these models. GMM estimators employ a set of population moment conditions to determine the unknown parameters of the models. The simplest stochastic volatility model is known as the log-normal SV model,

$$r_t = \sigma_t \varepsilon_t \quad (6.16)$$

$$\ln \sigma_t^2 = \omega + \rho \ln (\sigma_{t-1}^2 - \omega) + \sigma_\eta \eta_t \quad (6.17)$$

where $(\varepsilon_t, \eta_t) \stackrel{\text{i.i.d.}}{\sim} N(0, I_2)$ are i.i.d. standard normal. The first equation specifies the distribution of returns as heteroskedastic normal. The second equation specifies the dynamics of the log of volatility as an AR(1). The parameter vector is $(\omega, \rho, \sigma_\eta)'$. The application of GMM will use functions of r_t to identify the parameters of the model. Because this model is so simple, it is straight forward to derive the following relationships:

$$\begin{aligned} E[|r_t|] &= \sqrt{\frac{2}{\pi}} E[\sigma_t] \\ E[r_t^2] &= E[\sigma_t^2] \\ E[|r_t^3|] &= 2\sqrt{\frac{2}{\pi}} E[\sigma_t^3] \\ E[|r_t^4|] &= 3E[\sigma_t^4] \\ E[|r_t r_{t-j}|] &= \frac{2}{\pi} E[\sigma_t \sigma_{t-j}] \\ E[|r_t^2 r_{t-j}^2|] &= E[\sigma_t^2 \sigma_{t-j}^2] \end{aligned} \quad (6.18)$$

where

$$\begin{aligned} E[\sigma_t^m] &= \exp \left(m \frac{\omega}{2} + m^2 \frac{\sigma_\eta^2}{8} \right) \\ E[\sigma_t^m \sigma_{t-j}^n] &= E[\sigma_t^m] E[\sigma_{t-j}^n] \exp \left((mn) \rho^j \frac{\sigma_\eta^2}{4} \right). \end{aligned} \quad (6.19)$$

These conditions provide a large set of moments to determine the three unknown parameters. GMM seamlessly allows 3 or more moment conditions to be used in the estimation of the unknowns.

6.3 General Specification

The three examples show how a model – economic or statistical – can be turned into a set of moment conditional that have zero expectation, at least if the model is correctly specified. All GMM specifications are constructed this way. Derivation of GMM begins by defining the population moment condition.

Definition 6.1 (Population Moment Condition). Let \mathbf{w}_t be a vector of random variables, θ_0 be a p by 1 vector of parameters, and $\mathbf{g}(\cdot)$ be a q by 1 vector valued function. The population moment condition is defined

$$E[\mathbf{g}(\mathbf{w}_t, \theta_0)] = \mathbf{0} \quad (6.20)$$

It is possible that $\mathbf{g}(\cdot)$ could change over time and so could be replaced with $\mathbf{g}_t(\cdot)$. For clarity of exposition the more general case will not be considered.

Definition 6.2 (Sample Moment Condition). The sample moment condition is derived from the average population moment condition,

$$\mathbf{g}_T(\mathbf{w}, \theta) = T^{-1} \sum_{t=1}^T \mathbf{g}(\mathbf{w}_t, \theta). \quad (6.21)$$

The \mathbf{g}_T notation dates back to the original paper of Hansen (1982) and is widely used to differentiate population and sample moment conditions. Also note that the sample moment condition suppresses the t in \mathbf{w} . The GMM estimator is defined as the value of θ that minimizes

$$Q_T(\theta) = \mathbf{g}_T(\mathbf{w}, \theta)' \mathbf{W}_T \mathbf{g}_T(\mathbf{w}, \theta). \quad (6.22)$$

Thus the GMM estimator is defined as

$$\hat{\theta} = \arg \min_{\theta} Q_T(\theta) \quad (6.23)$$

where \mathbf{W}_T is a q by q positive semi-definite matrix. \mathbf{W}_T may (and generally will) depend on the data but it is required to converge in probability to a positive definite matrix for the estimator to be well defined. In order to operationalize the GMM estimator, q , the number of moments, will be required to greater than or equal to p , the number of unknown parameters.

6.3.1 Identification and Overidentification

GMM specifications fall in to three categories: underidentified, just-identified and overidentified. Underidentified models are those where the number of non-redundant moment conditions is less than the number of parameters. The consequence of this is obvious: the problem will have many solutions. Just-identified specification have $q = p$ while overidentified GMM specifications have $q > p$. The role of just- and overidentification will be reexamined in the context of estimation and inference. In most applications of GMM it is sufficient to count the number of moment equations and the number of parameters when determining whether the model is just- or overidentified. The exception to this rule arises if some moment conditions are linear combination of other moment conditions – in other words are redundant – which is similar to including a perfectly co-linear variable in a regression.

6.3.1.1 Example: Consumption Asset Pricing Model

In the consumption asset pricing model, the population moment condition is given by

$$\mathbf{g}(\mathbf{w}_t, \theta_0) = \left(\beta_0 (1 + \mathbf{r}_{t+1}) \left(\frac{c_{t+1}}{c_t} \right)^{-\gamma_0} - 1 \right) \otimes \mathbf{z}_t \quad (6.24)$$

where $\theta_0 = (\beta_0, \gamma_0)'$, and $\mathbf{w}_t = (c_{t+1}, c_t, \mathbf{r}'_{t+1}, \mathbf{z}'_t)'$ and \otimes denotes Kronecker product.² Note that both \mathbf{r}_{t+1} and \mathbf{z}_t are column vectors and so if there are n assets and k instruments, then the dimension of $\mathbf{g}(\cdot)$ (number of moment conditions) is $q = nk$ by 1 and the number of parameters is $p = 2$. Systems with $nk \geq 2$ will be identified as long as some technical conditions are met regarding *instrument validity* (see section 6.11).

6.3.1.2 Example: Linear Factor Models

In the linear factor models, the population moment conditions are given by

$$\mathbf{g}(\mathbf{w}_t, \theta_0) = \begin{pmatrix} (\mathbf{r}_t - \beta \mathbf{f}_t) \otimes \mathbf{f}_t \\ \mathbf{r}_t - \beta \lambda \end{pmatrix} \quad (6.27)$$

where $\theta_0 = (\text{vec}(\beta)', \lambda')'$ and $\mathbf{w}_t = (\mathbf{r}'_t, \mathbf{f}'_t)'$ where \mathbf{r}_t is n by 1 and \mathbf{f}_t is k by 1.³ These moments can be decomposed into two blocks. The top block contains the moment conditions necessary to estimate the β s. This block can be further decomposed into n blocks of k moment conditions, one for each factor. The first of these n blocks is

²

Definition 6.3 (Kronecker Product). Let $\mathbf{A} = [a_{ij}]$ be an m by n matrix, and let $\mathbf{B} = [b_{ij}]$ be a k by l matrix. The Kronecker product is defined

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \dots & a_{2n}\mathbf{B} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{bmatrix} \quad (6.25)$$

and has dimension mk by nl . If \mathbf{a} and \mathbf{b} are column vectors with length m and k respectively, then

$$\mathbf{a} \otimes \mathbf{b} = \begin{bmatrix} a_1\mathbf{b} \\ a_2\mathbf{b} \\ \vdots \\ a_m\mathbf{b} \end{bmatrix}. \quad (6.26)$$

³The vec operator stacks the columns of a matrix into a column vector.

Definition 6.4 (vec). Let $\mathbf{A} = [a_{ij}]$ be an m by n matrix. The the vec operator (also known as the *stack* operator) is defined

$$\text{vec}\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \\ \vdots \\ \mathbf{A}_n \end{bmatrix} \quad (6.28)$$

and $\text{vec}(\mathbf{A})$ is mn by 1.

$$\begin{bmatrix} (r_{1t} - \beta_{11}f_{1t} - \beta_{12}f_{2t} - \dots - \beta_{1K}f_{Kt})f_{1t} \\ (r_{1t} - \beta_{11}f_{1t} - \beta_{12}f_{2t} - \dots - \beta_{1K}f_{Kt})f_{2t} \\ \vdots \\ (r_{1t} - \beta_{11}f_{1t} - \beta_{12}f_{2t} - \dots - \beta_{1K}f_{Kt})f_{Kt} \end{bmatrix} = \begin{bmatrix} \varepsilon_{1t}f_{1t} \\ \varepsilon_{1t}f_{2t} \\ \vdots \\ \varepsilon_{1t}f_{Kt} \end{bmatrix}. \quad (6.29)$$

Each equation in (6.29) should be recognized as the first order condition for estimating the slope coefficients in a linear regression. The second block has the form

$$\begin{bmatrix} r_{1t} - \beta_{11}\lambda_1 - \beta_{12}\lambda_2 - \dots - \beta_{1K}\lambda_K \\ r_{2t} - \beta_{21}\lambda_1 - \beta_{22}\lambda_2 - \dots - \beta_{2K}\lambda_K \\ \vdots \\ r_{Nt} - \beta_{N1}\lambda_1 - \beta_{N2}\lambda_2 - \dots - \beta_{NK}\lambda_K \end{bmatrix} \quad (6.30)$$

where λ_j is the risk premium on the j^{th} factor. These moment conditions are derived from the relationship that the average return on an asset should be the sum of its risk exposure times the premium for that exposure.

The number of moment conditions (and the length of $\mathbf{g}(\cdot)$) is $q = nk + n$. The number of parameters is $p = nk$ (from β) + k (from λ), and so the number of overidentifying restrictions is the number of equations in $\mathbf{g}(\cdot)$ minus the number of parameters, $(nk + n) - (nk + k) = n - k$, the same number of restrictions used when testing asset pricing models in a two-stage Fama-MacBeth regressions.

6.3.1.3 Example: Stochastic Volatility Model

Many moment conditions are available to use in the stochastic volatility model. It is clear that at least 3 conditions are necessary to identify the 3 parameters and that the upper bound on the number of moment conditions is larger than the amount of data available. For clarity of exposition, only 5 and 8 moment conditions will be used, where the 8 are a superset of the 5. These 5 are:

$$\mathbf{g}(\mathbf{w}_t, \theta_0) = \begin{bmatrix} |r_t| - \sqrt{\frac{2}{\pi}} \exp\left(\frac{\omega}{2} + \frac{\sigma_\eta^2}{8}\right) \\ r_t^2 - \exp\left(\omega + \frac{\sigma_\eta^2}{2}\right) \\ r_t^4 - 3 \exp\left(2\omega + 2\frac{\sigma_\eta^2}{1-\rho^2}\right) \\ |r_t r_{t-1}| - \frac{2}{\pi} \left(\exp\left(\frac{\omega}{2} + \frac{\sigma_\eta^2}{8}\right) \right)^2 \exp\left(\rho \frac{\sigma_\eta^2}{4}\right) \\ r_t^2 r_{t-2}^2 - \left(\exp\left(\omega + \frac{\sigma_\eta^2}{2}\right) \right)^2 \exp\left(\rho^2 \frac{\sigma_\eta^2}{1-\rho^2}\right) \end{bmatrix} \quad (6.31)$$

These moment conditions can be easily verified from 6.18 and 6.19. The 8 moment-condition estimation extends the 5 moment-condition estimation with

$$\mathbf{g}(\mathbf{w}_t, \theta_0) = \begin{bmatrix} \text{Moment conditions from 6.31} \\ |r_t^3| - 2\sqrt{\frac{2}{\pi}} \exp\left(3\frac{\omega}{2} + 9\frac{\sigma_\eta^2}{8}\right) \\ |r_t r_{t-3}| - \frac{2}{\pi} \left(\exp\left(\frac{\omega}{2} + \frac{\sigma_\eta^2}{8}\right) \right)^2 \exp\left(\rho^3 \frac{\sigma_\eta^2}{4}\right) \\ r_t^2 r_{t-4}^2 - \left(\exp\left(\omega + \frac{1-\rho^2}{2}\right) \right)^2 \exp\left(\rho^4 \frac{\sigma_\eta^2}{1-\rho^2}\right) \end{bmatrix} \quad (6.32)$$

The moments that use lags are all staggered to improve identification of ρ .

6.4 Estimation

Estimation of GMM is seemingly simple but in practice fraught with difficulties and user choices. From the definitions of the GMM estimator,

$$Q_T(\theta) = \mathbf{g}_T(\mathbf{w}, \theta)' \mathbf{W}_T \mathbf{g}_T(\mathbf{w}, \theta) \quad (6.33)$$

$$\hat{\theta} = \arg \min_{\theta} Q_T(\theta) \quad (6.34)$$

Differentiation can be used to find the solution, $\hat{\theta}$, which solves

$$2\mathbf{G}_T(\mathbf{w}, \hat{\theta})' \mathbf{W}_T \mathbf{g}_T(\mathbf{w}, \hat{\theta}) = \mathbf{0} \quad (6.35)$$

where $\mathbf{G}_T(\mathbf{w}, \theta)$ is the q by p Jacobian of the moment conditions with respect to θ' ,

$$\mathbf{G}_T(\mathbf{w}, \theta) = \frac{\partial \mathbf{g}_T(\mathbf{w}, \theta)}{\partial \theta'} = T^{-1} \sum_{t=1}^T \frac{\partial \mathbf{g}(\mathbf{w}_t, \theta)}{\partial \theta'}. \quad (6.36)$$

$\mathbf{G}_T(\mathbf{w}, \theta)$ is a matrix of derivatives with q rows and p columns where each row contains the derivative of one of the moment conditions with respect to all p parameters and each column contains the derivative of the q moment conditions with respect to a single parameter.

The seeming simplicity of the calculus obscures two important points. First, the solution in eq. (6.35) does not generally emit an analytical solution and so numerical optimization must be used. Second, $Q_T(\cdot)$ is generally not a convex function in θ with a unique minimum, and so local minima are possible. The solution to the latter problem is to try multiple starting values and clever initial choices for starting values whenever available.

Note that \mathbf{W}_T has not been specified other than requiring that this weighting matrix is positive definite. The choice of the weighting matrix is an additional complication of using GMM. Theory dictates that the best choice of the weighting matrix must satisfy $\mathbf{W}_T \xrightarrow{p} \mathbf{S}^{-1}$ where

$$\mathbf{S} = \text{avar} \left\{ \sqrt{T} \mathbf{g}_T(\mathbf{w}_t, \theta_0) \right\} \quad (6.37)$$

and where *avar* indicates asymptotic variance. That is, the best choice of weighting is the inverse of the covariance of the moment conditions. Unfortunately the covariance of the moment conditions generally depends on the *unknown* parameter vector, θ_0 . The usual solution is to use multi-step estimation. In the first step, a simple choice for W_T , which does not depend on θ (often \mathbf{I}_q the identity matrix), is used to estimate $\hat{\theta}$. The second uses the first-step estimate of $\hat{\theta}$ to estimate $\hat{\mathbf{S}}$. A more formal discussion of the estimation of \mathbf{S} will come later. For now, assume that a consistent estimation method is being used so that $\hat{\mathbf{S}} \xrightarrow{P} \mathbf{S}$ and so $\mathbf{W}_T = \hat{\mathbf{S}}^{-1} \xrightarrow{P} \mathbf{S}^{-1}$.

The three main methods used to implement GMM are the classic 2-step estimation, K -step estimation where the estimation only ends after some convergence criteria is met and continuous updating estimation.

6.4.1 2-step Estimator

Two-step estimation is the standard procedure for estimating parameters using GMM. First-step estimates are constructed using a preliminary weighting matrix $\tilde{\mathbf{W}}$, often the identity matrix, and $\hat{\theta}_1$ solves the initial optimization problem

$$2\mathbf{G}_T(\mathbf{w}, \hat{\theta}_1)' \tilde{\mathbf{W}} \mathbf{g}_T(\mathbf{w}, \hat{\theta}_1) = \mathbf{0}. \quad (6.38)$$

The second step uses an estimated $\hat{\mathbf{S}}$ based on the first-step estimates $\hat{\theta}_1$. For example, if the moments are a martingale difference sequence with finite covariance,

$$\hat{\mathbf{S}}(\mathbf{w}, \hat{\theta}_1) = T^{-1} \sum_{t=1}^T \mathbf{g}(\mathbf{w}_t, \hat{\theta}_1) \mathbf{g}(\mathbf{w}_t, \hat{\theta}_1)' \quad (6.39)$$

is a consistent estimator of the asymptotic variance of $\mathbf{g}_T(\cdot)$, and the second-step estimates, $\hat{\theta}_2$, minimizes

$$Q_T(\theta) = \mathbf{g}_T(\mathbf{w}, \theta)' \hat{\mathbf{S}}^{-1}(\hat{\theta}_1) \mathbf{g}_T(\mathbf{w}, \theta). \quad (6.40)$$

which has first order condition

$$2\mathbf{G}_T(\mathbf{w}, \hat{\theta}_2)' \hat{\mathbf{S}}^{-1}(\hat{\theta}_1) \mathbf{g}_T(\mathbf{w}, \hat{\theta}_2) = \mathbf{0}. \quad (6.41)$$

Two-step estimation relies on the the consistence of the first-step estimates, $\hat{\theta}_1 \xrightarrow{P} \theta_0$ which is generally needed for $\hat{\mathbf{S}} \xrightarrow{P} \mathbf{S}$.

6.4.2 k -step Estimator

The k -step estimation strategy extends the two-step estimator in an obvious way: if two-steps are better than one, k may be better than two. The k -step procedure picks up where the 2-step procedure left off and continues iterating between $\hat{\theta}$ and $\hat{\mathbf{S}}$ using the most recent values $\hat{\theta}$ available when computing

the covariance of the moment conditions. The procedure terminates when some stopping criteria is satisfied. For example if

$$\max |\hat{\theta}_k - \hat{\theta}_{k-1}| < \varepsilon \quad (6.42)$$

for some small value ε , the iterations would stop and $\hat{\theta} = \hat{\theta}_k$. The stopping criteria should depend on the values of θ . For example, if these values are close to 1, then 1×10^{-4} may be a good choice for a stopping criteria. If the values are larger or smaller, the stopping criteria should be adjusted accordingly. The k -step and the 2-step estimator are asymptotically equivalent, although, the k -step procedure is thought to have better small sample properties than the 2-step estimator, particularly when it converges.

6.4.3 Continuously Updating Estimator (CUE)

The final, and most complicated, type of estimation, is the continuously updating estimator. Instead of iterating between estimation of θ and \mathbf{S} , this estimator parametrizes \mathbf{S} as a function of θ . In the problem, $\hat{\theta}_C$ is found as the minimum of

$$Q_T^C(\theta) = \mathbf{g}_T(\mathbf{w}, \theta)' \mathbf{S}(\mathbf{w}, \theta)^{-1} \mathbf{g}_T(\mathbf{w}, \theta) \quad (6.43)$$

The first order condition of this problem is *not* the same as in the original problem since θ appears in three terms. However, the estimates are still first-order asymptotically equivalent to the two-step estimate (and hence the k -step as well), and if the continuously updating estimator converges, it is generally regarded to have the best small sample properties among these methods.⁴ There are two caveats to using the continuously updating estimator. First, it is necessary to ensure that $\mathbf{g}_T(\mathbf{w}, \theta)$ is close to zero and that minimum is not being determined by a large covariance since a large $\mathbf{S}(\mathbf{w}, \theta)$ will make $Q_T^C(\theta)$ small for any value of the sample moment conditions $\mathbf{g}_T(\mathbf{w}, \theta)$. The second warning when using the continuously updating estimator has to make sure that $\mathbf{S}(\mathbf{w}, \theta)$ is not singular. If the weighting matrix is singular, there are values of θ which satisfy the first order condition which are not consistent. The continuously updating estimator is usually implemented using the k -step estimator to find starting values. Once the k -step has converged, switch to the continuously updating estimator until it also converges.

6.4.4 Improving the first step (when it matters)

There are two important caveats to the first-step choice of weighting matrix. The first is simple: if the problem is just identified, then the choice of weighting matrix does not matter and only one step is needed. To understand this, consider the first-order condition which defines $\hat{\theta}$,

$$2\mathbf{G}_T(\mathbf{w}, \hat{\theta})' \mathbf{W}_T \mathbf{g}_T(\mathbf{w}, \hat{\theta}) = \mathbf{0}. \quad (6.44)$$

If the number of moment conditions is the same as the number of parameters, the solution must have

$$\mathbf{g}_T(\mathbf{w}, \hat{\theta}) = \mathbf{0}. \quad (6.45)$$

⁴The continuously updating estimator is more efficient in the second-order sense than the 2- of k -step estimators, which improves finite sample properties.

as long as \mathbf{W}_T is positive definite and $\mathbf{G}_T(\mathbf{w}, \hat{\theta})$ has full rank (a necessary condition). However, if this is true, then

$$2\mathbf{G}_T(\mathbf{w}, \hat{\theta})'\tilde{\mathbf{W}}_T\mathbf{g}_T(\mathbf{w}, \hat{\theta}) = \mathbf{0} \quad (6.46)$$

for any other positive definite $\tilde{\mathbf{W}}_T$ whether it is the identity matrix, the asymptotic variance of the moment conditions, or something else.

The other important concern when choosing the initial weighting matrix is to not overweight high-variance moments and underweight low variance ones. Reasonable first-step estimates improve the estimation of $\hat{\mathbf{S}}$ which in turn provide more accurate second-step estimates. The second (and later) steps automatically correct for the amount of variability in the moment conditions. One fairly robust starting value is to use a diagonal matrix with the *inverse* of the variances of the moment conditions on the diagonal. This requires knowledge about θ to implement and an initial estimate or a good guess can be used. Asymptotically it makes no difference, although careful weighing in first-step estimation improves the performance of the 2-step estimator.

6.4.5 Example: Consumption Asset Pricing Model

The consumption asset pricing model example will be used to illustrate estimation. The data set consists of two return series, the value-weighted market portfolio and the equally-weighted market portfolio, *VWM* and *EWM* respectively. Models were fit to each return series separately. Real consumption data was available from Q1 1947 until Q4 2009 and downloaded from FRED (PCECC96). Five instruments (\mathbf{z}_t) will be used, a constant (1), contemporaneous and lagged consumption growth (c_t/c_{t-1} and c_{t-1}/c_{t-2}) and contemporaneous and lagged gross returns on the VWM (p_t/p_{t-1} and p_{t-1}/p_{t-2}). Using these five instruments, the model is overidentified since there are only 2 unknowns and five moment conditions,

$$\mathbf{g}(\mathbf{w}_t, \theta_0) = \begin{bmatrix} \left(\beta (1 + r_{t+1}) \left(\frac{c_{t+1}}{c_t} \right)^{-\gamma} - 1 \right) \\ \left(\beta (1 + r_{t+1}) \left(\frac{c_{t+1}}{c_t} \right)^{-\gamma} - 1 \right) \frac{c_t}{c_{t-1}} \\ \left(\beta (1 + r_{t+1}) \left(\frac{c_{t+1}}{c_t} \right)^{-\gamma} - 1 \right) \frac{c_{t-1}}{c_{t-2}} \\ \left(\beta (1 + r_{t+1}) \left(\frac{c_{t+1}}{c_t} \right)^{-\gamma} - 1 \right) \frac{p_t}{p_{t-1}} \\ \left(\beta (1 + r_{t+1}) \left(\frac{c_{t+1}}{c_t} \right)^{-\gamma} - 1 \right) \frac{p_{t-1}}{p_{t-2}} \end{bmatrix} \quad (6.47)$$

where r_{t+1} is the return on either the VWM or the EWM. Table 6.1 contains parameter estimates using the 4 methods outlined above for each asset.

The parameters estimates were broadly similar across the different estimators. The typical discount rate is very low (β close to 1) and the risk aversion parameter appears to be between 0.5 and 2.

One aspect of the estimation of this model is that γ is not well identified. Figure 6.1 contain surface and contour plots of the objective function as a function of β and γ for both the two-step estimator and the CUE. It is obvious in both pictures that the objective function is steep along the β -axis but very flat along the γ -axis. This means that γ is not well identified and many values will

Method	VWM		EWM	
	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\beta}$	$\hat{\gamma}$
Initial weighting matrix : \mathbf{I}_5				
1-Step	0.977	0.352	0.953	2.199
2-Step	0.975	0.499	0.965	1.025
k -Step	0.975	0.497	0.966	0.939
Continuous	0.976	0.502	0.966	0.936
Initial weighting matrix: $(\mathbf{z}'\mathbf{z})^{-1}$				
1-Step	0.975	0.587	0.955	1.978
2-Step	0.975	0.496	0.966	1.004
k -Step	0.975	0.497	0.966	0.939
Continuous	0.976	0.502	0.966	0.936

Table 6.1: Parameter estimates from the consumption asset pricing model using both the VWM and the EWM to construct the moment conditions. The top block corresponds to using an identity matrix for starting values while the bottom block of four correspond to using $(\mathbf{z}'\mathbf{z})^{-1}$ in the first step. The first-step estimates seem to be better behaved and closer to the 2- and K -step estimates when $(\mathbf{z}'\mathbf{z})^{-1}$ is used in the first step. The K -step and continuously updating estimators both converged and so produce the same estimates irrespective of the 1-step weighting matrix.

result in nearly the same objective function value. These results demonstrate how difficult GMM can be in even a simple 2-parameter model. Significant care should always be taken to ensure that the objective function has been globally minimized.

6.4.6 Example: Stochastic Volatility Model

The stochastic volatility model was fit using both 5 and 8 moment conditions to the returns on the FTSE 100 from January 1, 2000 until December 31, 2009, a total of 2,525 trading days. The results of the estimation are in table 6.2. The parameters differ substantially between the two methods. The 5-moment estimates indicate relatively low persistence of volatility with substantial variability. The 8-moment estimates all indicate that volatility is extremely persistent with ρ close to 1. All estimates weighting matrix computed using a Newey-West covariance with 16 lags ($\approx 1.2T^{\frac{1}{3}}$). A non-trivial covariance matrix is needed in this problem as the moment conditions should be persistent in the presence of stochastic volatility, unlike in the consumption asset pricing model which should, if correctly specified, have martingale errors.

In all cases the initial weighting matrix was specified to be an identity matrix, although in estimation problems such as this where the moment condition can be decomposed into $\mathbf{g}(\mathbf{w}_t, \theta) = \mathbf{f}(\mathbf{w}_t) - \mathbf{h}(\theta)$ a simple expression for the covariance can be derived by noting that, if the model is well specified, $E[\mathbf{g}(\mathbf{w}_t, \theta)] = \mathbf{0}$ and thus $\mathbf{h}(\theta) = E[\mathbf{f}(\mathbf{w}_t)]$. Using this relationship the covariance of $\mathbf{f}(\mathbf{w}_t)$ can be computed replacing $\mathbf{h}(\theta)$ with the sample mean of $\mathbf{f}(\mathbf{w}_t)$.

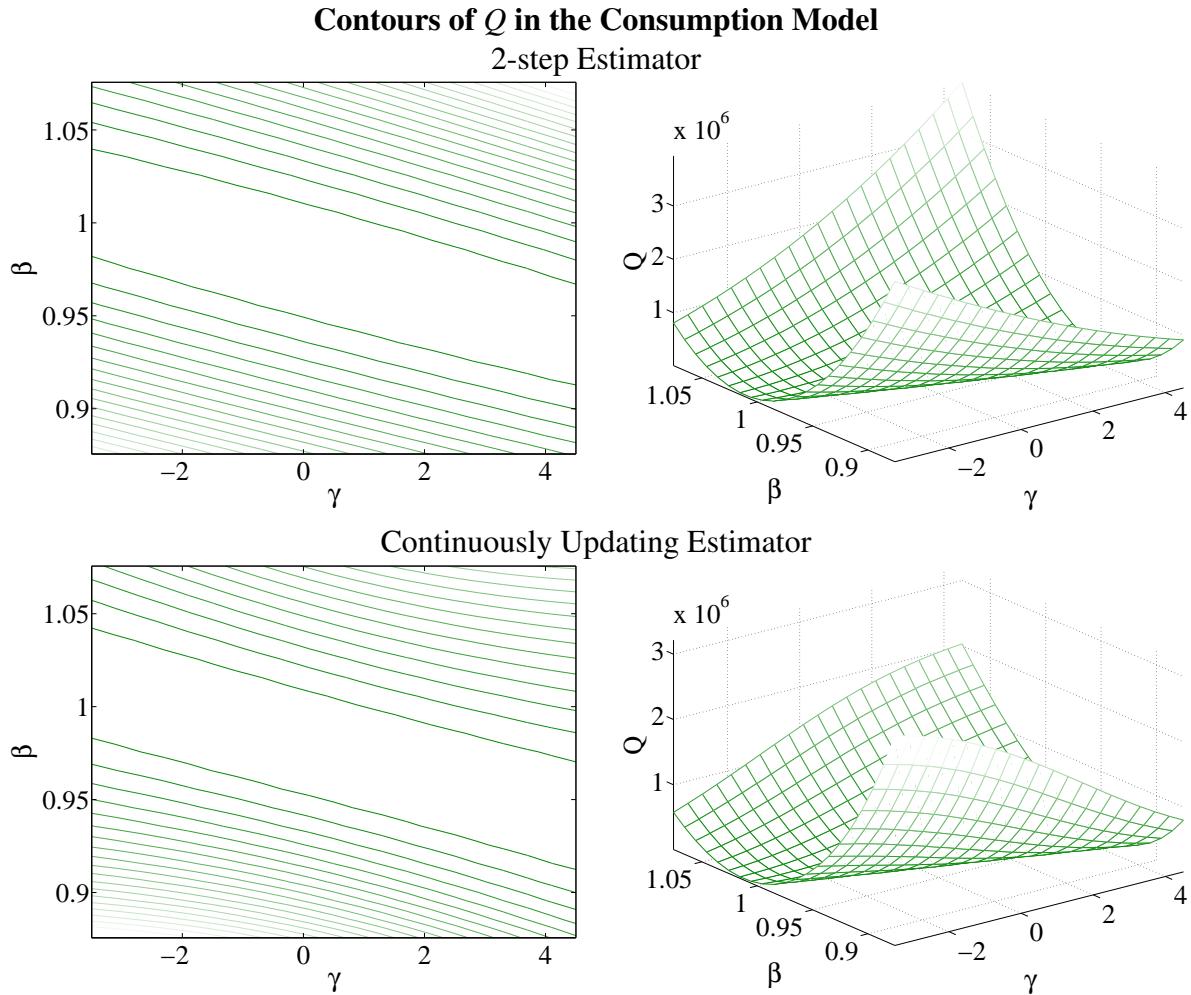


Figure 6.1: This figure contains a plot of the GMM objective function using the 2-step estimator (top panels) and the CUE (bottom panels). The objective is very steep along the β axis but nearly flat along the γ axis. This indicates that γ is not well identified.

6.5 Asymptotic Properties

The GMM estimator is consistent and asymptotically normal under fairly weak, albeit technical, assumptions. Rather than list 7-10 (depending on which setup is being used) hard to interpret assumptions, it is more useful to understand *why* the GMM estimator is consistent and asymptotically normal. The key to developing this intuition comes from understanding the that moment conditions used to define the estimator, $\mathbf{g}_T(\mathbf{w}, \theta)$, are simple averages which should have mean 0 when the population moment condition is true.

In order for the estimates to be reasonable, $\mathbf{g}(\mathbf{w}_t, \theta)$ need to be well behaved. One scenario where this occurs is when $\mathbf{g}(\mathbf{w}_t, \theta)$ is a stationary, ergodic sequence with a few moments. If this is true, only a few additional assumptions are needed to ensure $\hat{\theta}$ should be consistent and asymptotically normal. Specifically, \mathbf{W}_T must be positive definite and the system must be identified. Positive definiteness of \mathbf{W}_T is required to ensure that $Q_T(\theta)$ can only be minimized at one value – θ_0 . If \mathbf{W}_T were positive semi-definite or indefinite, many values would minimize the objective function. Identification is

Method	$\hat{\omega}$	$\hat{\rho}$	$\hat{\sigma}_\eta$
5 moment conditions			
1-Step	0.004	1.000	0.005
2-Step	-0.046	0.865	0.491
k -Step	-0.046	0.865	0.491
Continuous	-0.046	0.865	0.491
8 moment conditions			
1-Step	0.060	1.000	0.005
2-Step	-0.061	1.000	0.005
k -Step	-0.061	1.000	0.004
Continuous	-0.061	1.000	0.004

Table 6.2: Parameter estimates from the stochastic volatility model using both the 5- and 8-moment condition specifications on the returns from the FTSE from January 1, 2000 until December 31, 2009.

trickier, but generally requires that there is enough variation in the moment conditions to uniquely determine all of the parameters. Put more technically, the rank of $\mathbf{G} = \text{plim} \mathbf{G}_T(\mathbf{w}, \theta_0)$ must be weakly larger than the number of parameters in the model. Identification will be discussed in more detail in section 6.11. If these technical conditions are true, then the GMM estimator has standard properties.

6.5.1 Consistency

The estimator is consistent under relatively weak conditions. Formal consistency arguments involve showing that $Q_T(\theta)$ is suitably close to $E[Q_T(\theta)]$ in large samples so that the minimum of the sample objective function is close to the minimum of the population objective function. The most important requirement – and often the most difficult to verify – is that the parameters are uniquely identified which is equivalently to saying that there is only one value θ_0 for which $E[\mathbf{g}(\mathbf{w}_t, \theta)] = \mathbf{0}$. If this condition is true, and some more technical conditions hold, then

$$\hat{\theta} - \theta_0 \xrightarrow{P} 0 \quad (6.48)$$

The important point of this result is that the estimator is consistent for any choice of \mathbf{W}_T , not just $\mathbf{W}_T \xrightarrow{P} \mathbf{S}^{-1}$ since whenever \mathbf{W}_T is positive definite and the parameters are uniquely identified, $Q_T(\theta)$ can only be minimized when $E[\mathbf{g}(\mathbf{w}, \theta)] = 0$ which is θ_0 .

6.5.2 Asymptotic Normality of GMM Estimators

The GMM estimator is also asymptotically normal, although the form of the asymptotic covariance depends on how the parameters are estimated. Asymptotic normality of GMM estimators follows from taking a mean-value (similar to a Taylor) expansion of the moment conditions around the true parameter θ_0 ,

$$\begin{aligned} \mathbf{0} = \mathbf{G}_T(\mathbf{w}, \hat{\theta})' \mathbf{W}_T \mathbf{g}(\mathbf{w}, \hat{\theta}) &\approx \mathbf{G}' \mathbf{W} \mathbf{g}(\mathbf{w}, \theta_0) + \mathbf{G}' \mathbf{W} \frac{\partial \mathbf{g}(\mathbf{w}, \ddot{\theta})}{\partial \theta'} (\hat{\theta} - \theta_0) \\ &\approx \mathbf{G}' \mathbf{W} \mathbf{g}(\mathbf{w}, \theta_0) + \mathbf{G}' \mathbf{W} \mathbf{G} (\hat{\theta} - \theta_0) \end{aligned} \quad (6.49)$$

$$\begin{aligned} \mathbf{G}' \mathbf{W} \mathbf{G} (\hat{\theta} - \theta_0) &\approx -\mathbf{G}' \mathbf{W} \mathbf{g}(\mathbf{w}, \theta_0) \\ \sqrt{T} (\hat{\theta} - \theta_0) &\approx -(\mathbf{G}' \mathbf{W} \mathbf{G})^{-1} \mathbf{G}' \mathbf{W} [\sqrt{T} \mathbf{g}(\mathbf{w}, \theta_0)] \end{aligned} \quad (6.50)$$

where $\mathbf{G} \equiv \text{plim} \mathbf{G}_T(\mathbf{w}, \theta_0)$ and $\mathbf{W} \equiv \text{plim} \mathbf{W}_T$. The first line uses the score condition on the left hand side and the right-hand side contains the first-order Taylor expansion of the first-order condition. The second line uses the definition $\mathbf{G} = \partial \mathbf{g}(\mathbf{w}, \theta) / \partial \theta'$ evaluated at some point $\ddot{\theta}$ between $\hat{\theta}$ and θ_0 (element-by-element) the last line scales the estimator by \sqrt{T} . This expansion shows that the asymptotic normality of GMM estimators is derived directly from the normality of the moment conditions evaluated at the true parameter – moment conditions which are averages and so may, subject to some regularity conditions, follow a CLT.

The asymptotic variance of the parameters can be computed by computing the variance of the last line in eq. (6.49).

$$\begin{aligned} \text{V}[\sqrt{T} (\hat{\theta} - \theta_0)] &= (\mathbf{G}' \mathbf{W} \mathbf{G})^{-1} \mathbf{G}' \mathbf{W} \text{V}[\sqrt{T} \mathbf{g}(\mathbf{w}, \theta_0), \sqrt{T} \mathbf{g}(\mathbf{w}, \theta_0)'] \mathbf{W}' \mathbf{G} (\mathbf{G}' \mathbf{W} \mathbf{G})^{-1} \\ &= (\mathbf{G}' \mathbf{W} \mathbf{G})^{-1} \mathbf{G}' \mathbf{W} \mathbf{S} \mathbf{W}' \mathbf{G} (\mathbf{G}' \mathbf{W} \mathbf{G})^{-1} \end{aligned}$$

Using the asymptotic variance, the asymptotic distribution of the GMM estimator is

$$\sqrt{T} (\hat{\theta} - \theta_0) \xrightarrow{d} N(\mathbf{0}, (\mathbf{G}' \mathbf{W} \mathbf{G})^{-1} \mathbf{G}' \mathbf{W} \mathbf{S} \mathbf{W}' \mathbf{G} (\mathbf{G}' \mathbf{W} \mathbf{G})^{-1}) \quad (6.51)$$

If one were to use single-step estimation with an identity weighting matrix, the asymptotic covariance would be $(\mathbf{G}' \mathbf{G})^{-1} \mathbf{G}' \mathbf{S} \mathbf{G} (\mathbf{G}' \mathbf{G})^{-1}$. This format may look familiar: the White heteroskedasticity robust standard error formula when $\mathbf{G} = \mathbf{X}$ are the regressors and $\mathbf{G}' \mathbf{S} \mathbf{G} = \mathbf{X}' \mathbf{E} \mathbf{X}$, where \mathbf{E} is a diagonal matrix composed of the squared regression errors.

6.5.2.1 Asymptotic Normality, efficient \mathbf{W}

This form of the covariance simplifies when the efficient choice of $\mathbf{W} = \mathbf{S}^{-1}$ is used,

$$\begin{aligned} \text{V}[\sqrt{T} (\hat{\theta} - \theta_0)] &= (\mathbf{G}' \mathbf{S}^{-1} \mathbf{G})^{-1} \mathbf{G}' \mathbf{S}^{-1} \mathbf{S} \mathbf{S}^{-1} \mathbf{G} (\mathbf{G}' \mathbf{S}^{-1} \mathbf{G})^{-1} \\ &= (\mathbf{G}' \mathbf{S}^{-1} \mathbf{G})^{-1} \mathbf{G}' \mathbf{S}^{-1} \mathbf{G} (\mathbf{G}' \mathbf{S}^{-1} \mathbf{G})^{-1} \\ &= (\mathbf{G}' \mathbf{S}^{-1} \mathbf{G})^{-1} \end{aligned}$$

and the asymptotic distribution is

$$\sqrt{T} (\hat{\theta} - \theta_0) \xrightarrow{d} N(\mathbf{0}, (\mathbf{G}' \mathbf{S}^{-1} \mathbf{G})^{-1}) \quad (6.52)$$

Using the long-run variance of the moment conditions produces an asymptotic covariance which is not only simpler than the generic form, but is also *smaller* (in the matrix sense). This can be verified since

$$\begin{aligned} (\mathbf{G}'\mathbf{W}\mathbf{G})^{-1} \mathbf{G}'\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{G} (\mathbf{G}'\mathbf{W}\mathbf{G})^{-1} - (\mathbf{G}'\mathbf{S}^{-1}\mathbf{G})^{-1} &= \\ (\mathbf{G}'\mathbf{W}\mathbf{G})^{-1} \left[\mathbf{G}'\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{G} - (\mathbf{G}'\mathbf{W}\mathbf{G}) (\mathbf{G}'\mathbf{S}^{-1}\mathbf{G})^{-1} (\mathbf{G}'\mathbf{W}\mathbf{G}) \right] (\mathbf{G}'\mathbf{W}\mathbf{G})^{-1} &= \\ (\mathbf{G}'\mathbf{W}\mathbf{G})^{-1} \mathbf{G}'\mathbf{W}\mathbf{S}^{\frac{1}{2}} \left[\mathbf{I}_q - \mathbf{S}^{-\frac{1}{2}}\mathbf{G} (\mathbf{G}'\mathbf{S}^{-1}\mathbf{G})^{-1} \mathbf{G}'\mathbf{S}^{-\frac{1}{2}} \right] \mathbf{S}^{\frac{1}{2}}\mathbf{W}\mathbf{G} (\mathbf{G}'\mathbf{W}\mathbf{G})^{-1} &= \mathbf{A}' \left[\mathbf{I}_q - \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \right] \mathbf{A} \end{aligned}$$

where $\mathbf{A} = \mathbf{S}^{\frac{1}{2}}\mathbf{W}\mathbf{G} (\mathbf{G}'\mathbf{W}\mathbf{G})^{-1}$ and $\mathbf{X} = \mathbf{S}^{-\frac{1}{2}}\mathbf{G}$. This is a quadratic form where the inner matrix is idempotent – and hence positive semi-definite – and so the difference must be weakly positive. In most cases the efficient weighting matrix should be used, although there are application where an alternative choice of covariance matrix must be made due to practical considerations (many moment conditions) or for testing specific hypotheses.

6.5.3 Asymptotic Normality of the estimated moments, $\mathbf{g}_T(\mathbf{w}, \hat{\theta})$

Not only are the parameters asymptotically normal, but the estimated moment conditions are as well. The asymptotic normality of the moment conditions allows for testing the specification of the model by examining whether the sample moments are sufficiently close to 0. If the efficient weighting matrix is used ($\mathbf{W} = \mathbf{S}^{-1}$),

$$\sqrt{T}\mathbf{W}_T^{1/2}\mathbf{g}_T(\mathbf{w}, \hat{\theta}) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{I}_q - \mathbf{W}^{1/2}\mathbf{G} [\mathbf{G}'\mathbf{W}\mathbf{G}]^{-1} \mathbf{G}'\mathbf{W}^{1/2}\right) \quad (6.53)$$

The variance appears complicated but has a simple intuition. If the true parameter vector was known, $\mathbf{W}_T^{1/2}\hat{\mathbf{g}}_T(\mathbf{w}, \theta)$ would be asymptotically normal with identity covariance matrix. The second term is a result of having to estimate an unknown parameter vector. Essentially, one degree of freedom is lost for every parameter estimated and the covariance matrix of the estimated moments has $q - p$ degrees of freedom remaining. Replacing \mathbf{W} with the efficient choice of weighting matrix (\mathbf{S}^{-1}), the asymptotic variance of $\sqrt{T}\hat{\mathbf{g}}_T(\mathbf{w}, \hat{\theta})$ can be equivalently written as $\mathbf{S} - \mathbf{G} [\mathbf{G}'\mathbf{S}^{-1}\mathbf{G}]^{-1} \mathbf{G}'$ by pre- and post-multiplying the variance in by $\mathbf{S}^{\frac{1}{2}}$. In cases where the model is just-identified, $q = p$, $\mathbf{g}_T(\mathbf{w}, \hat{\theta}) = \mathbf{0}$, and the asymptotic variance is degenerate ($\mathbf{0}$).

If some other weighting matrix is used where $\mathbf{W}_T \xrightarrow{p} \mathbf{S}$ then the asymptotic distribution of the moment conditions is more complicated.

$$\sqrt{T}\mathbf{W}_T^{1/2}\mathbf{g}_T(\mathbf{w}, \hat{\theta}) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{N}\mathbf{W}^{1/2}\mathbf{S}\mathbf{W}^{1/2}\mathbf{N}'\right) \quad (6.54)$$

where $\mathbf{N} = \mathbf{I}_q - \mathbf{W}^{1/2}\mathbf{G} [\mathbf{G}'\mathbf{W}\mathbf{G}]^{-1} \mathbf{G}'\mathbf{W}^{1/2}$. If an alternative weighting matrix is used, the estimated moments are still asymptotically normal but with a different, *larger* variance. To see how the efficient form of the covariance matrix is nested in this inefficient form, replace $\mathbf{W} = \mathbf{S}^{-1}$ and note that since \mathbf{N} is idempotent, $\mathbf{N} = \mathbf{N}'$ and $\mathbf{N}\mathbf{N} = \mathbf{N}$.

6.6 Covariance Estimation

Estimation of the long run (asymptotic) covariance matrix of the moment conditions is important and often has significant impact on tests of either the model or individual coefficients. Recall the definition of the long-run covariance,

$$\mathbf{S} = \text{avar} \left\{ \sqrt{T} \mathbf{g}_T(\mathbf{w}_t, \theta_0) \right\}.$$

\mathbf{S} is the covariance of an average, $\mathbf{g}_T(\mathbf{w}_t, \theta_0) = \sqrt{T} T^{-1} \sum_{t=1}^T \mathbf{g}(\mathbf{w}_t, \theta_0)$ and the variance of an average includes all autocovariance terms. The simplest estimator one could construct to capture all autocovariance is

$$\hat{\mathbf{S}} = \hat{\Gamma}_0 + \sum_{i=1}^{T-1} \left(\hat{\Gamma}_i + \hat{\Gamma}'_i \right) \quad (6.55)$$

where

$$\hat{\Gamma}_i = T^{-1} \sum_{t=i+1}^T \mathbf{g}(\mathbf{w}_t, \hat{\theta}) \mathbf{g}(\mathbf{w}_{t-i}, \hat{\theta})'.$$

While this estimator is the natural sample analogue of the population long-run covariance, it is not positive definite and so is not useful in practice. A number of alternatives have been developed which can capture autocorrelation in the moment conditions and are guaranteed to be positive definite.

6.6.1 Serially uncorrelated moments

If the moments are serially uncorrelated then the usual covariance estimator can be used. Moreover, $E[\mathbf{g}_T(\mathbf{w}, \theta)] = 0$ and so \mathbf{S} can be consistently estimated using

$$\hat{\mathbf{S}} = T^{-1} \sum_{t=1}^T \mathbf{g}(\mathbf{w}_t, \hat{\theta}) \mathbf{g}(\mathbf{w}_t, \hat{\theta})', \quad (6.56)$$

This estimator does not explicitly remove the mean of the moment conditions. In practice it may be important to ensure the mean of the moment condition when the problem is over-identified ($q > p$), and is discussed further in [6.6.5](#).

6.6.2 Newey-West

The Newey and West (1987) covariance estimator solves the problem of positive definiteness of an autocovariance robust covariance estimator by weighting the autocovariances. This produces a heteroskedasticity, autocovariance consistent (HAC) covariance estimator that is guaranteed to be positive definite. The Newey-West estimator computes the long-run covariance as if the moment process was a vector moving average (VMA), and uses the sample autocovariances, and is defined (for a maximum lag l)

$$\hat{\mathbf{S}}^{NW} = \hat{\Gamma}_0 + \sum_{i=1}^l \frac{l+1-i}{l+1} (\hat{\Gamma}_i + \hat{\Gamma}'_i) \quad (6.57)$$

The number of lags, l , is problem dependent, and in general must grow with the sample size to ensure consistency when the moment conditions are dependent. The optimal rate of lag growth has $l = cT^{\frac{1}{3}}$ where c is a problem specific constant.

6.6.3 Vector Autoregressive

While the Newey-West covariance estimator is derived from a VMA, a Vector Autoregression (VAR)-based estimator can also be used. The VAR-based long-run covariance estimators have significant advantages when the moments are highly persistent. Construction of the VAR HAC estimator is simple and is derived directly from a VAR. Suppose the moment conditions, \mathbf{g}_t follow a VAR(r), and so

$$\mathbf{g}_t = \Phi_0 + \Phi_1 \mathbf{g}_{t-1} + \Phi_2 \mathbf{g}_{t-2} + \dots + \Phi_r \mathbf{g}_{t-r} + \eta_t. \quad (6.58)$$

The long run covariance of \mathbf{g}_t can be computed from knowledge of Φ_j , $j = 1, 2, \dots, s$ and the covariance of η_t . Moreover, if the assumption of VAR(r) dynamics is correct, η_t is a white noise process and its covariance can be consistently estimated by

$$\hat{\mathbf{S}}_\eta = (T - r)^{-1} \sum_{t=r+1}^T \hat{\eta}_t \hat{\eta}'_t. \quad (6.59)$$

The long run covariance is then estimated using

$$\hat{\mathbf{S}}^{AR} = (\mathbf{I} - \hat{\Phi}_1 - \hat{\Phi}_2 - \dots - \hat{\Phi}_r)^{-1} \hat{\mathbf{S}}_\eta \left((\mathbf{I} - \hat{\Phi}_1 - \hat{\Phi}_2 - \dots - \hat{\Phi}_r)^{-1} \right)'. \quad (6.60)$$

The primary advantage of the VAR based estimator over the NW is that the number of parameters needing to be estimated is often much, much smaller. If the process is well described by an VAR, k may be as small as 1 while a Newey-West estimator may require many lags to adequately capture the dependence in the moment conditions. Haan and Levin (2000) show that the VAR procedure can be consistent if the number of lags grow as the sample size grows so that the VAR can approximate the long-run covariance of any covariance stationary process. They recommend choosing the lag length using BIC in two steps: first choosing the lag length of own lags, and then choosing the number of lags of other moments.

6.6.4 Pre-whitening and Recoloring

The Newey-West and VAR long-run covariance estimators can be combined in a procedure known as pre-whitening and recoloring. This combination exploits the VAR to capture the persistence in the moments and used the Newey-West HAC to capture any neglected serial dependence in the residuals. The advantage of this procedure over Newey-West or VAR HAC covariance estimators is that PWRC is parsimonious while allowing complex dependence in the moments.

A low order VAR (usually 1st) is fit to the moment conditions,

$$\mathbf{g}_t = \Phi_0 + \Phi_1 \mathbf{g}_{t-1} + \eta_t \quad (6.61)$$

and the covariance of the residuals, $\hat{\eta}_t$ is estimated using a Newey-West estimator, preferably with a small number of lags,

$$\hat{\mathbf{S}}_{\eta}^{NW} = \hat{\Xi}_0 + \sum_{i=1}^l \frac{l-i+1}{l+1} (\hat{\Xi}_i + \hat{\Xi}'_i) \quad (6.62)$$

where

$$\hat{\Xi}_i = T^{-1} \sum_{t=i+1}^T \hat{\eta}_t \hat{\eta}'_{t-i}. \quad (6.63)$$

The long run covariance is computed by combining the VAR parameters with the Newey-West covariance of the residuals,

$$\hat{\mathbf{S}}^{PWRC} = (\mathbf{I} - \hat{\Phi}_1)^{-1} \hat{\mathbf{S}}_{\eta}^{NW} ((\mathbf{I} - \hat{\Phi}_1)^{-1})', \quad (6.64)$$

or, if a higher order VAR was used,

$$\hat{\mathbf{S}}^{PWRC} = \left(\mathbf{I} - \sum_{j=1}^r \hat{\Phi}_j \right)^{-1} \hat{\mathbf{S}}_{\eta}^{NW} \left(\left(\mathbf{I} - \sum_{j=1}^r \hat{\Phi}_j \right)^{-1} \right)' \quad (6.65)$$

where r is the order of the VAR.

6.6.5 To demean or not to demean?

One important issue when computing asymptotic variances is whether the sample moments should be demeaned before estimating the long-run covariance. If the population moment conditions are valid, then $E[\mathbf{g}_t(\mathbf{w}_t, \theta)] = \mathbf{0}$ and the covariance can be computed from $\{\mathbf{g}_t(\mathbf{w}_t, \hat{\theta})\}$ without removing the mean. If the population moment conditions are *not* valid, then $E[\mathbf{g}_t(\mathbf{w}_t, \theta)] \neq \mathbf{0}$ and any covariance matrices estimated from the sample moments will be inconsistent. The intuition behind the inconsistency is simple. Suppose the $E[\mathbf{g}_t(\mathbf{w}_t, \theta)] \neq \mathbf{0}$ for all $\theta \in \Theta$, the parameter space and that the moments are a vector martingale process. Using the “raw” moments to estimate the covariance produces an inconsistent estimator since

$$\hat{\mathbf{S}} = T^{-1} \sum_{t=1}^T \mathbf{g}(\mathbf{w}_t, \hat{\theta}) \mathbf{g}(\mathbf{w}_t, \hat{\theta})' \xrightarrow{P} \mathbf{S} + \mu \mu' \quad (6.66)$$

where \mathbf{S} is the covariance of the moment conditions and μ is the expectation of the moment conditions evaluated at the probability limit of the first-step estimator, $\hat{\theta}_1$.

One way to remove the inconsistency is to demean the moment conditions prior to estimating the long run covariance so that $\mathbf{g}_t(\mathbf{w}_t, \hat{\theta})$ is replaced by $\tilde{\mathbf{g}}_t(\mathbf{w}_t, \hat{\theta}) = \mathbf{g}_t(\mathbf{w}_t, \hat{\theta}) - T^{-1} \sum_{t=1}^T \mathbf{g}_t(\mathbf{w}_t, \hat{\theta})$ when computing the long-run covariance. Note that demeaning is not *free* since removing the mean, when the population moment condition is valid, reduces the variation in $\mathbf{g}_t(\cdot)$ and in turn the precision of

\hat{S} . As a general rule, the mean should be removed except in cases where the sample length is small relative to the number of moments. In practice, subtracting the mean from the estimated moment conditions is important for testing models using J -tests and for estimating the parameters in 2- or k -step procedures.

6.6.6 Example: Consumption Asset Pricing Model

Returning to the consumption asset pricing example, 11 different estimators were used to estimate the long run variance after using the parameters estimated in the first step of the GMM estimator. These estimators include the standard estimator and both the Newey-West and the VAR estimator using 1 to 5 lags. In this example, the well identified parameter, β is barely affected but the poorly identified γ shows some variation when the covariance estimator changes. In this example it is reasonable to use the simple covariance estimator because, if the model is well specified, the moments *must* be serially uncorrelated. If they are serially correlated then the investor's marginal utility is predictable and so the model is misspecified. It is generally good practice to impose any theoretically sound restrictions on the covariance estimator (such as a lack of serial correlation in this example, or at most some finite order moving average).

Lags	Newey-West		Autoregressive	
	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\beta}$	$\hat{\gamma}$
0	0.975	0.499		
1	0.979	0.173	0.982	-0.082
2	0.979	0.160	0.978	0.204
3	0.979	0.200	0.977	0.399
4	0.978	0.257	0.976	0.493
5	0.978	0.276	0.976	0.453

Table 6.3: The choice of variance estimator can have an effect on the estimated parameters in a 2-step GMM procedure. The estimate of the discount rate is fairly stable, but the estimate of the coefficient of risk aversion changes as the long-run covariance estimator varies. Note that the NW estimation with 0 lags is the just the usual covariance estimator.

6.6.7 Example: Stochastic Volatility Model

Unlike the previous example, efficient estimation in the stochastic volatility model example *requires* the use of a HAC covariance estimator. The stochastic volatility estimator uses unconditional moments which are serially correlated whenever the data has time-varying volatility. For example, the moment conditions $E[|r_t|]$ is autocorrelated since $E[|r_t r_{t-j}|] \neq E[|r_t|]^2$ (see eq.(6.19)). All of the parameter estimates in table 6.2 were computed using a Newey-West covariance estimator with 12 lags, which was chosen using $cT^{\frac{1}{3}}$ rule where $c = 1.2$ was chosen. Rather than use actual data to investigate the value of various HAC estimators, consider a simple Monte Carlo where the DGP is

$$\begin{aligned} r_t &= \sigma_t \varepsilon_t \\ \ln \sigma_t^2 &= -7.36 + 0.9 \ln (\sigma_{t-1}^2 - 7.36) + 0.363 \eta_t \end{aligned} \tag{6.67}$$

which corresponds to an annualized volatility of 22%. Both shocks were standard normal. 1000 replications with $T = 1000$ and 2500 were conducted and the covariance matrix was estimated using 4 different estimators: a misspecified covariance assuming the moments are uncorrelated, a HAC using $1.2T^{1/3}$, a VAR estimator where the lag length is automatically chosen by the SIC, and an “infeasible” estimate computed using a Newey-West estimator computed from an auxiliary run of 10 million simulated data points. The first-step estimator was estimated using an identity matrix.

The results of this small simulation study are presented in table 6.4. This table contains a lot of information, much of which is contradictory. It highlights the difficulties in actually making the correct choices when implementing a GMM estimator. For example, the bias of the 8 moment estimator is generally at least as large as the bias from the 5 moment estimator, although the root mean square error is generally better. This highlights the general bias-variance trade-off that is made when using more moments: more moments leads to less variance but more bias. The only absolute rule evident from the the table is the performance changes when moving from 5 to 8 moment conditions and using the *infeasible* covariance estimator. The additional moments contained information about both the persistence ρ and the volatility of volatility σ .

6.7 Special Cases of GMM

GMM can be viewed as a unifying class which nests mean estimators. Estimators used in frequentist econometrics can be classified into one of three types: M-estimators (maximum), R-estimators (rank), and L-estimators (linear combination). Most estimators presented in this course, including OLS and MLE, are in the class of M-estimators. All M-class estimators are the solution to some extremum problem such as minimizing the sum of squares or maximizing the log likelihood.

In contrast, all R-estimators make use of rank statistics. The most common examples include the minimum, the maximum and rank correlation, a statistic computed by calculating the usual correlation on the rank of the data rather than on the data itself. R-estimators are robust to certain issues and are particularly useful in analyzing nonlinear relationships. L-estimators are defined as any linear combination of rank estimators. The classical example of an L-estimator is a trimmed mean, which is similar to the usual mean estimator except some fraction of the data in each tail is eliminated, for example the top and bottom 1%. L-estimators are often substantially more robust than their M-estimator counterparts and often only make small sacrifices in efficiency. Despite the potential advantages of L-estimators, strong assumptions are needed to justify their use and difficulties in deriving theoretical properties limit their practical application.

GMM is obviously an M-estimator since it is the result of a minimization and any estimator nested in GMM must also be an M-estimator and most M-estimators are nested in GMM. The most important exception is a subclass of estimators known as classical minimum distance (CMD). CMD estimators minimize the distance between a restricted parameter space and an initial set of estimates. The final parameter estimates generally includes fewer parameters than the initial estimate or non-linear restrictions. CMD estimators are not widely used in financial econometrics, although they occasionally allow for feasible solutions to otherwise infeasible problems – usually because direct

	5 moment conditions			8 moment conditions		
	Bias					
T=1000	ω	ρ	σ	ω	ρ	σ
Inefficeint	-0.000	-0.024	-0.031	0.001	-0.009	-0.023
Serial Uncorr.	0.013	0.004	-0.119	0.013	0.042	-0.188
Newey-West	-0.033	-0.030	-0.064	-0.064	-0.009	-0.086
VAR	-0.035	-0.038	-0.058	-0.094	-0.042	-0.050
Infeasible	-0.002	-0.023	-0.047	-0.001	-0.019	-0.015
 T=2500						
Inefficeint	0.021	-0.017	-0.036	0.021	-0.010	-0.005
Serial Uncorr.	0.027	-0.008	-0.073	0.030	0.027	-0.118
Newey-West	-0.001	-0.034	-0.027	-0.022	-0.018	-0.029
VAR	0.002	-0.041	-0.014	-0.035	-0.027	-0.017
Infeasible	0.020	-0.027	-0.011	0.020	-0.015	0.012
 RMSE						
T=1000						
Inefficeint	0.121	0.127	0.212	0.121	0.073	0.152
Serial Uncorr.	0.126	0.108	0.240	0.128	0.081	0.250
Newey-West	0.131	0.139	0.217	0.141	0.082	0.170
VAR	0.130	0.147	0.218	0.159	0.132	0.152
Infeasible	0.123	0.129	0.230	0.128	0.116	0.148
 T=2500						
Inefficeint	0.075	0.106	0.194	0.075	0.055	0.114
Serial Uncorr.	0.079	0.095	0.201	0.082	0.065	0.180
Newey-West	0.074	0.102	0.182	0.080	0.057	0.094
VAR	0.072	0.103	0.174	0.085	0.062	0.093
Infeasible	0.075	0.098	0.185	0.076	0.054	0.100

Table 6.4: Results from the Monte Carlo experiment on the SV model. Two data lengths ($T = 1000$ and $T = 2500$) and two sets of moments were used. The table shows how difficult it can be to find reliable rules for improving finite sample performance. The only clean gains come from increasing the sample size and/or number of moments.

estimation of the parameters in the restricted parameter space is difficult or impossible using nonlinear optimizers.

6.7.1 Classical Method of Moments

The obvious example of GMM is classical method of moments. Consider using MM to estimate the parameters of a normal distribution. The two estimators are

$$\mu = T^{-1} \sum_{t=1}^T y_t \quad (6.68)$$

$$\sigma^2 = T^{-1} \sum_{t=1}^T (y_t - \mu)^2 \quad (6.69)$$

which can be transformed into moment conditions

$$\mathbf{g}_T(\mathbf{w}, \theta) = \begin{bmatrix} T^{-1} \sum_{t=1}^T y_t - \mu \\ T^{-1} \sum_{t=1}^T (y_t - \mu)^2 - \sigma^2 \end{bmatrix} \quad (6.70)$$

which obviously have the same solutions. If the data are i.i.d., then, defining $\hat{\epsilon}_t = y_t - \hat{\mu}$, \mathbf{S} can be consistently estimated by

$$\begin{aligned} \hat{\mathbf{S}} &= T^{-1} \sum_{t=1}^T [\mathbf{g}_t(\mathbf{w}_t, \hat{\theta}) \mathbf{g}_t(\mathbf{w}_t, \hat{\theta})'] \\ &= T^{-1} \sum_{t=1}^T \begin{bmatrix} \hat{\epsilon}_t^2 & \hat{\epsilon}_t (\hat{\epsilon}_t^2 - \sigma^2) \\ \hat{\epsilon}_t (\hat{\epsilon}_t^2 - \sigma^2) & (\hat{\epsilon}_t^2 - \sigma^2)^2 \end{bmatrix} \\ &= \begin{bmatrix} \sum_{t=1}^T \hat{\epsilon}_t^2 & \sum_{t=1}^T \hat{\epsilon}_t^3 \\ \sum_{t=1}^T \hat{\epsilon}_t^3 & \sum_{t=1}^T \hat{\epsilon}_t^4 - 2\sigma^2 \hat{\epsilon}_t^2 + \sigma^4 \end{bmatrix} \quad \text{since } \sum_{t=1}^T \hat{\epsilon}_t = 0 \\ \mathbb{E}[\hat{\mathbf{S}}] &\approx \begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix} \quad \text{if normal} \end{aligned} \quad (6.71)$$

Note that the last line is exactly the variance of the mean and variance if the covariance was estimated assuming normal maximum likelihood. Similarly, \mathbf{G}_T can be consistently estimated by

$$\begin{aligned} \hat{\mathbf{G}} &= T^{-1} \left[\begin{bmatrix} \frac{\partial \sum_{t=1}^T y_t - \mu}{\partial \mu} & \frac{\partial \sum_{t=1}^T (y_t - \mu)^2 - \sigma^2}{\partial \mu} \\ \frac{\partial \sum_{t=1}^T y_t - \mu}{\partial \sigma^2} & \frac{\partial \sum_{t=1}^T (y_t - \mu)^2 - \sigma^2}{\partial \sigma^2} \end{bmatrix} \right]_{\theta=\hat{\theta}} \\ &= T^{-1} \begin{bmatrix} \sum_{t=1}^T -1 & -2 \sum_{t=1}^T \hat{\epsilon}_t \\ 0 & \sum_{t=1}^T -1 \end{bmatrix} \\ &= T^{-1} \begin{bmatrix} \sum_{t=1}^T -1 & 0 \\ 0 & \sum_{t=1}^T -1 \end{bmatrix} \quad \text{by } \sum_{t=1}^T \hat{\epsilon}_t = 0 \end{aligned} \quad (6.72)$$

$$= T^{-1} \begin{bmatrix} -T & 0 \\ 0 & -T \end{bmatrix} \\ = -\mathbf{I}_2$$

and since the model is just-identified (as many moment conditions as parameters) $(\hat{\mathbf{G}}_T' \hat{\mathbf{S}}^{-1} \hat{\mathbf{G}}_T)^{-1} = (\hat{\mathbf{G}}_T^{-1})' \hat{\mathbf{S}} \hat{\mathbf{G}}_T^{-1} = \hat{\mathbf{S}}$, the usual covariance estimator for the mean and variance in the method of moments problem.

6.7.2 OLS

OLS (and other least squares estimators, such as WLS) can also be viewed as a special case of GMM by using the orthogonality conditions as the moments.

$$\mathbf{g}_T(\mathbf{w}, \theta) = T^{-1} \mathbf{X}' (\mathbf{y} - \mathbf{X}\beta) = T^{-1} \mathbf{X}' \boldsymbol{\varepsilon} \quad (6.73)$$

and the solution is obviously given by

$$\hat{\beta} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}. \quad (6.74)$$

If the data are from a stationary martingale difference sequence, then \mathbf{S} can be consistently estimated by

$$\begin{aligned} \hat{\mathbf{S}} &= T^{-1} \sum_{t=1}^T \mathbf{x}_t' \hat{\boldsymbol{\varepsilon}}_t \hat{\boldsymbol{\varepsilon}}_t \mathbf{x}_t \\ \hat{\mathbf{S}} &= T^{-1} \sum_{t=1}^T \hat{\boldsymbol{\varepsilon}}_t^2 \mathbf{x}_t' \mathbf{x}_t \end{aligned} \quad (6.75)$$

and \mathbf{G}_T can be estimated by

$$\begin{aligned} \hat{\mathbf{G}} &= T^{-1} \frac{\partial \mathbf{X}' (\mathbf{y} - \mathbf{X}\beta)}{\partial \beta'} \\ &= -T^{-1} \mathbf{X}' \mathbf{X} \end{aligned} \quad (6.76)$$

Combining these two, the covariance of the OLS estimator is

$$\left((-T^{-1} \mathbf{X}' \mathbf{X})^{-1} \right)' \left(T^{-1} \sum_{t=1}^T \hat{\boldsymbol{\varepsilon}}_t^2 \mathbf{x}_t' \mathbf{x}_t \right) (-T^{-1} \mathbf{X}' \mathbf{X})^{-1} = \hat{\Sigma}_{\mathbf{xx}}^{-1} \hat{\mathbf{S}} \hat{\Sigma}_{\mathbf{xx}}^{-1} \quad (6.77)$$

which is the White heteroskedasticity robust covariance estimator.

6.7.3 MLE and Quasi-MLE

GMM also nests maximum likelihood and quasi-MLE (QMLE) estimators. An estimator is said to be a QMLE if one distribution is assumed, for example normal, when the data are generated by some other distribution, for example a standardized Student's t . Most ARCH-type estimators are treated as QMLE since normal maximum likelihood is often used when the standardized residuals are clearly not normal, exhibiting both skewness and excess kurtosis. The most important consequence of QMLE is that the information matrix inequality is generally not valid and robust standard errors must be used. To formulate the (Q)MLE problem, the moment conditions are simply the average scores,

$$\mathbf{g}_T(\mathbf{w}, \theta) = T^{-1} \sum_{t=1}^T \nabla_{\theta} l(\mathbf{w}_t, \theta) \quad (6.78)$$

where $l(\cdot)$ is the log-likelihood. If the scores are a martingale difference sequence, \mathbf{S} can be consistently estimated by

$$\hat{\mathbf{S}} = T^{-1} \sum_{t=1}^T \nabla_{\theta} l(\mathbf{w}_t, \theta) \nabla_{\theta'} l(\mathbf{w}_t, \theta) \quad (6.79)$$

and \mathbf{G}_T can be estimated by

$$\hat{\mathbf{G}} = T^{-1} \sum_{t=1}^T \nabla_{\theta\theta'} l(\mathbf{w}_t, \theta). \quad (6.80)$$

However, in terms of expressions common to MLE estimation,

$$\begin{aligned} \mathbb{E}[\hat{\mathbf{S}}] &= \mathbb{E}[T^{-1} \sum_{t=1}^T \nabla_{\theta} l(\mathbf{w}_t, \theta) \nabla_{\theta'} l(\mathbf{w}_t, \theta)] \\ &= T^{-1} \sum_{t=1}^T \mathbb{E}[\nabla_{\theta} l(\mathbf{w}_t, \theta) \nabla_{\theta'} l(\mathbf{w}_t, \theta)] \\ &= T^{-1} T \mathbb{E}[\nabla_{\theta} l(\mathbf{w}_t, \theta) \nabla_{\theta'} l(\mathbf{w}_t, \theta)] \\ &= \mathbb{E}[\nabla_{\theta} l(\mathbf{w}_t, \theta) \nabla_{\theta'} l(\mathbf{w}_t, \theta)] \\ &= \mathcal{J} \end{aligned} \quad (6.81)$$

and

$$\begin{aligned} \mathbb{E}[\hat{\mathbf{G}}] &= \mathbb{E}[T^{-1} \sum_{t=1}^T \nabla_{\theta\theta'} l(\mathbf{w}_t, \theta)] \\ &= T^{-1} \sum_{t=1}^T \mathbb{E}[\nabla_{\theta\theta'} l(\mathbf{w}_t, \theta)] \\ &= T^{-1} T \mathbb{E}[\nabla_{\theta\theta'} l(\mathbf{w}_t, \theta)] \\ &= \mathbb{E}[\nabla_{\theta\theta'} l(\mathbf{w}_t, \theta)] \end{aligned} \quad (6.82)$$

$$\begin{aligned} &= E[\nabla_{\theta\theta'} l(\mathbf{w}_t, \theta)] \\ &= -\mathcal{I} \end{aligned}$$

The GMM covariance is $(\hat{\mathbf{G}}^{-1})' \hat{\mathbf{S}} \hat{\mathbf{G}}^{-1}$ which, in terms of MLE notation is $\mathcal{I}^{-1} \mathcal{J} \mathcal{I}^{-1}$. If the information matrix equality is valid ($\mathcal{I} = \mathcal{J}$), this simplifies to \mathcal{I}^{-1} , the usual variance in MLE. However, when the assumptions of the MLE procedure are not valid, the robust form of the covariance estimator, $\mathcal{I}^{-1} \mathcal{J} \mathcal{I}^{-1} = (\hat{\mathbf{G}}^{-1})' \hat{\mathbf{S}} \hat{\mathbf{G}}^{-1}$ should be used, and failure to do so can result in tests with incorrect size.

6.8 Diagnostics

The estimation of a GMM model begins by specifying the population moment conditions which, if correct, have mean 0. This is an assumption and is often a hypothesis of interest. For example, in the consumption asset pricing model the discounted returns should have conditional mean 0 and deviations from 0 indicate that the model is misspecified. The standard method to test whether the moment conditions is known as the J test and is defined

$$J = T \mathbf{g}_T(\mathbf{w}, \hat{\theta})' \hat{\mathbf{S}}^{-1} \mathbf{g}_T(\mathbf{w}, \hat{\theta}) \quad (6.83)$$

$$= T Q_T(\hat{\theta}) \quad (6.84)$$

which is T times the minimized GMM objective function where $\hat{\mathbf{S}}$ is a consistent estimator of the long-run covariance of the moment conditions. The distribution of J is χ^2_{q-p} , where $q-p$ measures the degree of overidentification. The distribution of the test follows directly from the asymptotic normality of the estimated moment conditions (see section 6.5.3). It is important to note that the standard J test requires the use of a multi-step estimator which uses an efficient weighting matrix ($\mathbf{W}_T \xrightarrow{P} \mathbf{S}^{-1}$).

In cases where an efficient estimator of \mathbf{W}_T is not available, an inefficient test test can be computed using

$$\begin{aligned} J^{W_T} &= T \mathbf{g}_T(\mathbf{w}, \hat{\theta})' \left(\left[\mathbf{I}_q - \mathbf{W}^{1/2} \mathbf{G} [\mathbf{G}' \mathbf{W} \mathbf{G}]^{-1} \mathbf{G}' \mathbf{W}^{1/2} \right] \mathbf{W}^{1/2} \right. \\ &\quad \times \left. \mathbf{S} \mathbf{W}^{1/2} \left[\mathbf{I}_q - \mathbf{W}^{1/2} \mathbf{G} [\mathbf{G}' \mathbf{W} \mathbf{G}]^{-1} \mathbf{G}' \mathbf{W}^{1/2} \right]' \right)^{-1} \mathbf{g}_T(\mathbf{w}, \hat{\theta}) \end{aligned} \quad (6.85)$$

which follow directly from the asymptotic normality of the estimated moment conditions even when the weighting matrix is sub-optimally chosen. J^{W_T} , like J , is distributed χ^2_{q-p} , although it is *not* T times the first-step GMM objective. Note that the inverse in eq. (6.85) is of a reduced rank matrix and must be computed using a Moore-Penrose generalized inverse.

6.8.1 Example: Linear Factor Models

The CAPM will be used to examine the use of diagnostic tests. The CAPM was estimated on the 25 Fama-French 5 by 5 sort on size and BE/ME using data from 1926 until 2010. The moments in this specification can be described

Method	CAPM		3 Factor	
	$J \sim \chi^2_{24}$	p-val	$J \sim \chi^2_{22}$	p-val
2-Step	98.0	0.000	93.3	0.000
k -Step	98.0	0.000	92.9	0.000
Continuous	98.0	0.000	79.5	0.000
2-step NW	110.4	0.000	108.5	0.000
2-step VAR	103.7	0.000	107.8	0.000

Table 6.5: Values of the J test using different estimation strategies. All of the tests agree, although the continuously updating version is substantially smaller in the 3 factor model (but highly significant since distributed χ^2_{22}).

$$\mathbf{g}_t(\mathbf{w}_t, \boldsymbol{\theta}) = \begin{bmatrix} (\mathbf{r}_t^e - \beta \mathbf{f}_t) \otimes \mathbf{f}_t \\ (\mathbf{r}_t^e - \beta \lambda) \end{bmatrix} \quad (6.86)$$

where \mathbf{f}_t is the excess return on the market portfolio and \mathbf{r}_t^e is a 25 by 1 vector of excess returns on the FF portfolios. There are 50 moment conditions and 26 unknown parameters so this system is overidentified and the J statistic is χ^2_{24} distributed. The J -statistics were computed for the four estimation strategies previously described, the inefficient 1-step test, 2-step, K -step and continuously updating. The values of these statistics, contained in table 6.5, indicate the CAPM is overwhelmingly rejected. While only the simple covariance estimator was used to estimate the long run covariance, all of the moments are portfolio returns and this choice seems reasonable considering the lack of predictability of monthly returns. The model was then extended to include the size and momentum factors, which resulted in 100 moment equations and 78 (75β s + 3 risk premia) parameters, and so the J statistic is distributed as a χ^2_{22} .

6.9 Parameter Inference

6.9.1 The delta method and nonlinear hypotheses

Thus far, all hypothesis tests encountered have been linear, and so can be expressed $H_0 : \mathbf{R}\boldsymbol{\theta} - \mathbf{r} = \mathbf{0}$ where \mathbf{R} is a M by P matrix of linear restriction and \mathbf{r} is a M by 1 vector of constants. While linear restrictions are the most common type of null hypothesis, some interesting problems require tests of nonlinear restrictions.

Define $\mathbf{R}(\boldsymbol{\theta})$ to be a M by 1 vector valued *function*. From this nonlinear function, a nonlinear hypothesis can be specified $H_0 : \mathbf{R}(\boldsymbol{\theta}) = \mathbf{0}$. To test this hypothesis, the distribution of $\mathbf{R}(\boldsymbol{\theta})$ needs to be determined (as always, under the null). The **delta method** can be used to simplify finding this distribution in cases where $\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is asymptotically normal as long as $\mathbf{R}(\boldsymbol{\theta})$ is a continuously differentiable function of $\boldsymbol{\theta}$ at $\boldsymbol{\theta}_0$.

Definition 6.5 (Delta method). Let $\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \Sigma)$ where Σ is a positive definite covariance matrix. Further, suppose that $\mathbf{R}(\boldsymbol{\theta})$ is a continuously differentiable function of $\boldsymbol{\theta}$ from $\mathbb{R}^p \rightarrow \mathbb{R}^m$, $m \leq p$. Then,

$$\sqrt{T}(\mathbf{R}(\hat{\boldsymbol{\theta}}) - \mathbf{R}(\boldsymbol{\theta}_0)) \xrightarrow{d} N\left(0, \frac{\partial \mathbf{R}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'} \Sigma \frac{\partial \mathbf{R}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right) \quad (6.87)$$

This result is easy to relate to the class of linear restrictions, $\mathbf{R}(\theta) = \mathbf{R}\theta - \mathbf{r}$. In this class,

$$\frac{\partial \mathbf{R}(\theta_0)}{\partial \theta'} = \mathbf{R} \quad (6.88)$$

and the distribution under the null is

$$\sqrt{T}(\mathbf{R}\hat{\theta} - \mathbf{R}\theta_0) \xrightarrow{d} N(0, \mathbf{R}\Sigma\mathbf{R}') . \quad (6.89)$$

Once the distribution of the nonlinear function θ has been determined, using the delta method to conduct a nonlinear hypothesis test is straight forward with one big catch. The null hypothesis is $H_0 : \mathbf{R}(\theta_0) = 0$ and a Wald test can be calculated

$$W = T\mathbf{R}(\hat{\theta})' \left[\frac{\partial \mathbf{R}(\theta_0)}{\partial \theta'} \Sigma \frac{\partial \mathbf{R}(\theta_0)}{\partial \theta'} \right]^{-1} \mathbf{R}(\hat{\theta}). \quad (6.90)$$

The distribution of the Wald test is determined by the rank of $\frac{\mathbf{R}(\theta)}{\partial \theta'}$ evaluated under H_0 . In some simple cases the rank is obvious. For example, in the linear hypothesis testing framework, the rank of $\frac{\mathbf{R}(\theta)}{\partial \theta'}$ is simply the rank of the matrix \mathbf{R} . In a test of a hypothesis $H_0 : \theta_1\theta_2 - 1 = 0$,

$$\frac{\mathbf{R}(\theta)}{\partial \theta'} = \begin{bmatrix} \theta_2 \\ \theta_1 \end{bmatrix} \quad (6.91)$$

assuming there are two parameters in θ and the rank of $\frac{\mathbf{R}(\theta)}{\partial \theta'}$ must be one if the null is true since both parameters must be non-zero to have a product of 1. The distribution of a Wald test of this null is a χ^2_1 . However, consider a test of the null $H_0 : \theta_1\theta_2 = 0$. The Jacobian of this function is identical but the slight change in the null has large consequences. For this null to be true, one of three things much occur: $\theta_1 = 0$ and $\theta_2 \neq 0$, $\theta_1 \neq 0$ and $\theta_2 = 0$ or $\theta_1 = 0$ and $\theta_2 = 0$. In the first two cases, the rank of $\frac{\mathbf{R}(\theta)}{\partial \theta'}$ is 1. However, in the last case the rank is 0. When the rank of the Jacobian can take multiple values depending on the value of the true parameter, the distribution under the null is nonstandard and none of the standard tests are directly applicable.

6.9.2 Wald Tests

Wald tests in GMM are essentially identical to Wald tests in OLS; W is T times the standardized, summed and squared deviations from the null. If the efficient choice of \mathbf{W}_T is used,

$$\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} N(\mathbf{0}, (\mathbf{G}'\mathbf{S}^{-1}\mathbf{G})^{-1}) \quad (6.92)$$

and a Wald test of the (linear) null $H_0 : \mathbf{R}\theta - \mathbf{r} = 0$ is computed

$$W = T(\mathbf{R}\hat{\theta} - \mathbf{r})' \left[\mathbf{R}(\mathbf{G}'\mathbf{S}^{-1}\mathbf{G})^{-1} \mathbf{R}' \right]^{-1} (\mathbf{R}\hat{\theta} - \mathbf{r}) \xrightarrow{d} \chi_m^2 \quad (6.93)$$

where m is the rank of \mathbf{R} . Nonlinear hypotheses can be tested in an analogous manner using the delta method. When using the delta method, m is the rank of $\frac{\partial \mathbf{R}(\theta_0)}{\partial \theta'}$. If an inefficient choice of \mathbf{W}_T is used,

$$\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} N(\mathbf{0}, (\mathbf{G}'\mathbf{W}\mathbf{G})^{-1} \mathbf{G}'\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{G} (\mathbf{G}'\mathbf{W}\mathbf{G})^{-1}) \quad (6.94)$$

and

$$W = T(\mathbf{R}\hat{\theta} - \mathbf{r})' \mathbf{V}^{-1} (\mathbf{R}\hat{\theta} - \mathbf{r}) \xrightarrow{d} \chi_m^2 \quad (6.95)$$

where $\mathbf{V} = \mathbf{R}(\mathbf{G}'\mathbf{W}\mathbf{G})^{-1} \mathbf{G}'\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{G}(\mathbf{G}'\mathbf{W}\mathbf{G})^{-1} \mathbf{R}'$.

T-tests and t-stats are also valid and can be computed in the usual manner for single hypotheses,

$$t = \frac{\mathbf{R}\hat{\theta} - r}{\sqrt{\mathbf{V}}} \xrightarrow{d} N(0, 1) \quad (6.96)$$

where the form of \mathbf{V} depends on whether an efficient or inefficient choice of \mathbf{W}_T was used. In the case of the t-stat of a parameter,

$$t = \frac{\hat{\theta}_i}{\sqrt{\mathbf{V}_{[ii]}}} \xrightarrow{d} N(0, 1) \quad (6.97)$$

where $\mathbf{V}_{[ii]}$ indicates the element in the i^{th} diagonal position.

6.9.3 Likelihood Ratio (LR) Tests

Likelihood Ratio-like tests, despite GMM making no distributional assumptions, are available. Let $\hat{\theta}$ indicate the unrestricted parameter estimate and let $\tilde{\theta}$ indicate the solution of

$$\begin{aligned} \tilde{\theta} &= \arg \min_{\theta} Q_T(\theta) \\ &\text{subject to } \mathbf{R}\theta - \mathbf{r} = \mathbf{0} \end{aligned} \quad (6.98)$$

where $Q_T(\theta) = \mathbf{g}_T(\mathbf{w}, \theta)' \hat{\mathbf{S}}^{-1} \mathbf{g}_T(\mathbf{w}, \theta)$ and $\hat{\mathbf{S}}$ is an estimate of the long-run covariance of the moment conditions computed from the unrestricted model (using $\hat{\theta}$). A LR-like test statistic can be formed

$$LR = T(\mathbf{g}_T(\mathbf{w}, \tilde{\theta})' \hat{\mathbf{S}}^{-1} \mathbf{g}_T(\mathbf{w}, \tilde{\theta}) - \mathbf{g}_T(\mathbf{w}, \hat{\theta})' \hat{\mathbf{S}}^{-1} \mathbf{g}_T(\mathbf{w}, \hat{\theta})) \xrightarrow{d} \chi_m^2 \quad (6.99)$$

Implementation of this test has one *crucial* aspect. The covariance matrix of the moments used in the second-step estimation *must* be the same for the two models. Using different covariance estimates can produce a statistic which is not χ^2 .

The likelihood ratio-like test has one significant advantage: it is invariant to equivalent reparameterization of either the moment conditions or the restriction (if nonlinear) while the Wald test is not. The intuition behind this result is simple; LR-like tests will be constructed using the same values of Q_T for any equivalent reparameterization and so the numerical value of the test statistic will be unchanged.

6.9.4 LM Tests

LM tests are also available and are the result of solving the Lagrangian

$$\tilde{\theta} = \arg \min_{\theta} Q_T(\theta) - \lambda'(\mathbf{R}\theta - \mathbf{r}) \quad (6.100)$$

In the GMM context, LM tests examine how much larger the restricted moment conditions are than their unrestricted counterparts. The derivation is messy and computation is harder than either Wald or LR, but the form of the LM test statistic is

$$LM = T \mathbf{g}_T(\mathbf{w}, \tilde{\theta})' \mathbf{S}^{-1} \mathbf{G} (\mathbf{G}' \mathbf{S}^{-1} \mathbf{G})^{-1} \mathbf{G}' \mathbf{S}^{-1} \mathbf{g}_T(\mathbf{w}, \tilde{\theta}) \xrightarrow{d} \chi_m^2 \quad (6.101)$$

The primary advantage of the LM test is that it only requires estimation under the null which can, in some circumstances, be much simpler than estimation under the alternative. You should note that the number of moment conditions must be the same in the restricted model as in the unrestricted model.

6.10 Two-Stage Estimation

Many common problems involve the estimation of parameters in stages. The most common example in finance are Fama-MacBeth regressions(Fama and MacBeth, 1973) which use two sets of regressions to estimate the factor loadings and risk premia. Another example is models which first fit conditional variances and then, conditional on the conditional variances, estimate conditional correlations. To understand the impact of first-stage estimation error on second-stage parameters, it is necessary to introduce some additional notation to distinguish the first-stage moment conditions from the second stage moment conditions. Let $\mathbf{g}_{1T}(\mathbf{w}, \psi) = T^{-1} \sum_{t=1}^T \mathbf{g}_1(\mathbf{w}_t, \psi)$ and $\mathbf{g}_{2T}(\mathbf{w}, \psi, \theta) = T^{-1} \sum_{t=1}^T \mathbf{g}_2(\mathbf{w}_t, \psi, \theta)$ be the first- and second-stage moment conditions. The first-stage moment conditions only depend on a subset of the parameters, ψ , and the second-stage moment conditions depend on both ψ and θ . The first-stage moments will be used to estimate ψ and the second-stage moments will treat $\hat{\psi}$ as known when estimating θ . Assuming that both stages are just-identified, which is the usual scenario, then

$$\begin{aligned} \sqrt{T} \begin{bmatrix} \hat{\psi} - \psi \\ \hat{\theta} - \theta \end{bmatrix} &\xrightarrow{d} N \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, (\mathbf{G}^{-1})' \mathbf{S} \mathbf{G}^{-1} \right) \\ \mathbf{G} &= \begin{bmatrix} \mathbf{G}_{1\psi} & \mathbf{G}_{2\psi} \\ \mathbf{0} & \mathbf{G}_{2\theta} \end{bmatrix} \\ \mathbf{G}_{1\psi} &= \frac{\partial \mathbf{g}_{1T}}{\partial \psi'}, \quad \mathbf{G}_{2\psi} = \frac{\partial \mathbf{g}_{2T}}{\partial \psi'}, \quad \mathbf{G}_{2\theta} = \frac{\partial \mathbf{g}_{2T}}{\partial \theta'} \\ \mathbf{S} &= \text{avar} \left(\left[\sqrt{T} \mathbf{g}'_{1T}, \sqrt{T} \mathbf{g}'_{2T} \right]' \right) \end{aligned}$$

Application of the partitioned inverse shows that the asymptotic variance of the first-stage parameters is identical to the usual expression, and so $\sqrt{T} (\hat{\psi} - \psi) \xrightarrow{d} N(\mathbf{0}, \mathbf{G}_{1\psi}^{-1} \mathbf{S}_{\psi\psi} \mathbf{G}_{1\psi}^{-1})$ where $\mathbf{S}_{\psi\psi}$ is the upper block of \mathbf{S} which corresponds to the \mathbf{g}_1 moments. The distribution of the second-stage parameters differs from what would be found if the estimation of ψ was ignored, and so

$$\begin{aligned} \sqrt{T} (\hat{\theta} - \theta) &\xrightarrow{d} N \left(\mathbf{0}, \mathbf{G}_{2\theta}^{-1} \left[\left[-\mathbf{G}_{2\psi} \mathbf{G}_{1\psi}^{-1}, \mathbf{I} \right] \mathbf{S} \left[-\mathbf{G}_{2\psi} \mathbf{G}_{1\psi}^{-1}, \mathbf{I} \right]' \right] \mathbf{G}_{2\theta}^{-1} \right) \quad (6.102) \\ &= N \left(\mathbf{0}, \mathbf{G}_{2\theta}^{-1} \left[\mathbf{S}_{\theta\theta} - \mathbf{G}_{2\psi} \mathbf{G}_{1\psi}^{-1} \mathbf{S}_{\psi\theta} - \mathbf{S}_{\theta\psi} \mathbf{G}_{1\psi}^{-1} \mathbf{G}_{2\psi} + \mathbf{G}_{2\psi} \mathbf{G}_{1\psi}^{-1} \mathbf{S}_{\psi\psi} \mathbf{G}_{1\psi}^{-1} \mathbf{G}'_{2\psi} \right] \mathbf{G}_{2\theta}^{-1} \right). \end{aligned}$$

The intuition for this form comes from considering an expansion of the second stage moments first around the second-stage parameters, and the accounting for the additional variation due to the first-stage parameter estimates. Expanding the second-stage moments around the true second stage-parameters,

$$\sqrt{T}(\hat{\theta} - \theta) \approx -\mathbf{G}_{2\theta}^{-1}\sqrt{T}\mathbf{g}_{2T}(\mathbf{w}, \hat{\psi}, \theta_0).$$

If ψ were known, then this would be sufficient to construct the asymptotic variance. When ψ is estimated, it is necessary to expand the final term around the first-stage parameters, and so

$$\sqrt{T}(\hat{\theta} - \theta) \approx -\mathbf{G}_{2\theta}^{-1} \left[\sqrt{T}\mathbf{g}_{2T}(\mathbf{w}, \psi_0, \theta_0) + \mathbf{G}_{2\psi}\sqrt{T}(\hat{\psi} - \psi) \right]$$

which shows that the error in the estimation of ψ appears in the estimation error of θ . Finally, using the relationship $\sqrt{T}(\hat{\psi} - \psi) \approx -\mathbf{G}_{1\psi}^{-1}\sqrt{T}\mathbf{g}_{1T}(\mathbf{w}, \psi_0)$, the expression can be completed, and

$$\begin{aligned} \sqrt{T}(\hat{\theta} - \theta) &\approx -\mathbf{G}_{2\theta}^{-1} \left[\sqrt{T}\mathbf{g}_{2T}(\mathbf{w}, \psi_0, \theta_0) - \mathbf{G}_{2\psi}\mathbf{G}_{1\psi}^{-1}\sqrt{T}\mathbf{g}_{1T}(\mathbf{w}, \psi_0) \right] \\ &= -\mathbf{G}_{2\theta}^{-1} \left[\left[-\mathbf{G}_{2\psi}\mathbf{G}_{1\psi}^{-1}, \mathbf{I} \right] \sqrt{T} \begin{bmatrix} \mathbf{g}_{1T}(\mathbf{w}, \psi_0) \\ \mathbf{g}_{2T}(\mathbf{w}, \psi_0, \theta_0) \end{bmatrix} \right]. \end{aligned}$$

Squaring this expression and replacing the outer-product of the moment conditions with the asymptotic variance produces eq. (6.102).

6.10.1 Example: Fama-MacBeth Regression

Fama-MacBeth regression is a two-step estimation procedure where the first step is just-identified and the second is over-identified. The first-stage moments are used to estimate the factor loadings (β s) and the second-stage moments are used to estimate the risk premia. In an application to n portfolios and k factors there are $q_1 = nk$ moments in the first-stage,

$$\mathbf{g}_{1t}(\mathbf{w}_t \theta) = (\mathbf{r}_t - \beta \mathbf{f}_t) \otimes \mathbf{f}_t$$

which are used to estimate nk parameters. The second stage uses k moments to estimate k risk premia using

$$\mathbf{g}_{2t}(\mathbf{w}_t \theta) = \beta'(\mathbf{r}_t - \beta \lambda) .$$

It is necessary to account for the uncertainty in the estimation of β when constructing confidence intervals for λ . Correct inference can be made by estimating the components of eq. (6.102),

$$\begin{aligned} \hat{\mathbf{G}}_{1\beta} &= T^{-1} \sum_{t=1}^T \mathbf{I}_n \otimes \mathbf{f}_t \mathbf{f}'_t, \\ \hat{\mathbf{G}}_{2\beta} &= T^{-1} \sum_{t=1}^T (\mathbf{r}_t - \hat{\beta} \hat{\lambda})' \otimes \mathbf{I}_k - \hat{\beta}' \otimes \hat{\lambda}', \end{aligned}$$

	Correct			OLS - White			Correct			OLS - White	
	$\hat{\lambda}$	s.e.	t-stat	s.e.	t-stat		$\hat{\lambda}$	s.e.	t-stat	s.e.	t-stat
VWM ^e	7.987	2.322	3.440	0.643	12.417		6.508	2.103	3.095	0.812	8.013
SMB	–						2.843	1.579	1.800	1.651	1.722
HML	–						3.226	1.687	1.912	2.000	1.613

Table 6.6: Annual risk premia, correct and OLS - White standard errors from the CAPM and the Fama-French 3 factor mode.

$$\hat{\mathbf{G}}_{2\lambda} = T^{-1} \sum_{t=1}^T \hat{\beta}' \hat{\beta},$$

$$\hat{\mathbf{S}} = T^{-1} \sum_{t=1}^T \begin{bmatrix} (\mathbf{r}_t - \hat{\beta} \mathbf{f}_t) \otimes \mathbf{f}_t \\ \hat{\beta}' (\mathbf{r}_t - \hat{\beta} \hat{\lambda}) \end{bmatrix} \begin{bmatrix} (\mathbf{r}_t - \hat{\beta} \mathbf{f}_t) \otimes \mathbf{f}_t \\ \hat{\beta}' (\mathbf{r}_t - \hat{\beta} \hat{\lambda}) \end{bmatrix}'.$$

These expressions were applied to the 25 Fama-French size and book-to-market sorted portfolios. Table 6.6 contains the standard errors and t-stats which are computed using both incorrect inference – White standard errors which come from a standard OLS of the mean excess return on the β s – and the consistent standard errors which are computed using the expressions above. The standard error and t-stats for the excess return on the market change substantially when the parameter estimation error in β is included.

6.11 Weak Identification

The topic of **weak identification** has been a unifying theme in recent work on GMM and related estimations. Three types of identification have previously been described: underidentified, just-identified and overidentified. Weak identification bridges the gap between just-identified and underidentified models. In essence, a model is weakly identified if it is identified in a finite sample, but the amount of information available to estimate the parameters does not increase with the sample. This is a difficult concept, so consider it in the context of the two models presented in this chapter.

In the consumption asset pricing model, the moment conditions are all derived from

$$\left(\beta (1 + r_{j,t+1}) \left(\frac{c_{t+1}}{c_t} \right)^{-\gamma} - 1 \right) z_t. \quad (6.103)$$

Weak identification can appear in at least two places in this moment conditions. First, assume that $\frac{c_{t+1}}{c_t} \approx 1$. If it were exactly 1, then γ would be unidentified. In practice consumption is very smooth and so the variation in this ratio from 1 is small. If the variation is decreasing over time, this problem would be weakly identified. Alternatively suppose that the instrument used, z_t , is not related to future marginal utilities or returns at all. For example, suppose z_t is a simulated a random variable. In this case,

$$E \left[\left(\beta (1 + r_{j,t+1}) \left(\frac{c_{t+1}}{c_t} \right)^{-\gamma} - 1 \right) z_t \right] = E \left[\left(\beta (1 + r_{j,t+1}) \left(\frac{c_{t+1}}{c_t} \right)^{-\gamma} - 1 \right) \right] E[z_t] = 0 \quad (6.104)$$

for any values of the parameters and so the moment condition is always satisfied. The choice of instrument matters a great deal and should be made in the context of economic and financial theories.

In the example of the linear factor models, weak identification can occur if a factor which is not important for any of the included portfolios is used in the model. Consider the moment conditions,

$$\mathbf{g}(\mathbf{w}_t, \theta_0) = \begin{pmatrix} (\mathbf{r}_t - \beta \mathbf{f}_t) \otimes \mathbf{f}_t \\ \mathbf{r} - \beta \lambda \end{pmatrix}. \quad (6.105)$$

If one of the factors is totally unrelated to asset returns and has no explanatory power, all β s corresponding to that factor will limit to 0. However, if this occurs then the second set of moment conditions will be valid for any choice of λ_i ; λ_i is weakly identified. Weak identification will make most inference nonstandard and so the limiting distributions of most statistics are substantially more complicated. Unfortunately there are few easy fixes for this problem and common sense and economic theory must be used when examining many problems using GMM.

6.12 Considerations for using GMM

This chapter has provided a introduction to GMM. However, before applying GMM to every econometric problem, there are a few issues which should be considered.

6.12.1 The degree of overidentification

Overidentification is beneficial since it allows models to be tested in a simple manner using the J test. However, like most things in econometrics, there are trade-off when deciding how overidentified a model should be. Increasing the degree of overidentification by adding extra moments but not adding more parameters can lead to substantial small sample bias and poorly behaving tests. Adding extra moments also increases the dimension of the estimated long run covariance matrix, $\hat{\mathbf{S}}$ which can lead to size distortion in hypothesis tests. Ultimately, the number of moment conditions should be traded off against the sample size. For example, in linear factor model with n portfolios and k factors there are $n - k$ overidentifying restrictions and $nk + k$ parameters. If testing the CAPM with monthly data back to WWII (approx 700 monthly observations), the total number of moments should be kept under 150. If using quarterly data (approx 250 quarters), the number of moment conditions should be substantially smaller.

6.12.2 Estimation of the long run covariance

Estimation of the long run covariance is one of the most difficult issues when implementing GMM. Best practices are to to use the simplest estimator consistent with the data or theoretical restrictions which is usually the estimator with the smallest parameter count. If the moments can be reasonably assumed to be a martingale difference series then a simple outer-product based estimator is sufficient. HAC estimators should be avoided if the moments are not autocorrelated (or cross-correlated). If the

moments are persistent with geometrically decaying autocorrelation, a simple VAR(1) model may be enough.

Longer Exercises

Exercise 6.1. Suppose you were interested in testing a multi-factor model with 4 factors and excess returns on 10 portfolios.

1. How many moment conditions are there?
2. What are the moment conditions needed to estimate this model?
3. How would you test whether the model correctly prices all assets. What are you really testing?
4. What are the requirements for identification?
5. What happens if you include a factor that is not relevant to the returns of any series?

Exercise 6.2. Suppose you were interested in estimating the CAPM with (potentially) non-zero α s on the excess returns of two portfolios, r_1^e and r_2^e .

1. Describe the moment equations you would use to estimate the 4 parameters.
2. Is this problem underidentified, just-identified, or overidentified?
3. Describe how you would conduct a joint test of the null $H_0 : \alpha_1 = \alpha_2 = 0$ against an alternative that at least one was non-zero using a Wald test.
4. Describe how you would conduct a joint test of the null $H_0 : \alpha_1 = \alpha_2 = 0$ against an alternative that at least one was non-zero using a LR-like test.
5. Describe how you would conduct a joint test of the null $H_0 : \alpha_1 = \alpha_2 = 0$ against an alternative that at least one was non-zero using an LM test.

In all of the questions involving tests, you should explain all steps from parameter estimation to the final rejection decision.

Chapter 7

Univariate Volatility Modeling

Alternative references for volatility modeling include chapters 10 and 11 in Taylor (2005), chapter 21 of Hamilton (1994), and chapter 4 of Enders (2004). Many of the original articles have been collected in Engle (1995).

Engle (1982) introduced the ARCH model and, in doing so, modern financial econometrics. Measuring and modeling conditional volatility is the cornerstone of the field. Models used for analyzing conditional volatility can be extended to capture a variety of related phenomena including Value-at-Risk, Expected Shortfall, forecasting the complete density of financial returns and duration analysis. This chapter begins by examining the meaning of “volatility” - it has many - before turning attention to the ARCH-family of models. The chapter details estimation, inference, model selection, forecasting, and diagnostic testing. The chapter concludes by covering new methods of measuring volatility: *realized volatility*, which makes use of using ultra-high-frequency data, and *implied volatility*, a measure of volatility computed from options prices.

Volatility measurement and modeling is the foundation of financial econometrics. This chapter begins by introducing volatility as a meaningful concept and then describes a widely used framework for volatility analysis: the ARCH model. The chapter describes the most widely used members of the ARCH family, fundamental properties of each, estimation, inference and model selection. Attention then turns to a new tool in the measurement and modeling of financial volatility, *realized volatility*, before concluding with a discussion of option-based *implied volatility*.

7.1 Why does volatility change?

Time-varying volatility is a pervasive empirical regularity in financial time series, and it is difficult to find an asset return series which does *not* exhibit time-varying volatility. This chapter focuses on providing a statistical description of the time-variation of volatility but does not go into depth on the economic causes of time-varying volatility. Many explanations have been proffered to explain this phenomenon, and treated individually; none provide a complete characterization of the variation in volatility observed in financial returns.

- *News Announcements:* The arrival of unanticipated news (or “news surprises”) forces agents to update beliefs. These new beliefs lead to portfolio rebalancing and high volatility correspond to

periods when agents are incorporating the news and dynamically solving for new asset prices. While certain classes of assets have been shown to react to surprises, in particular, government bonds and foreign exchange, many appear to be unaffected by large surprises (see, *inter alia* Engle and Li (1998) and Andersen, Bollerslev, Diebold, and Vega (2007)). Additionally, news-induced periods of high volatility are generally short, often on the magnitude of 5 to 30-minutes and the apparent resolution of uncertainty is far too quick to explain the time-variation of volatility seen in asset prices.

- *Leverage*: When a firm is financed using both debt and equity, only the equity reflects the volatility of the firm's cash flows. However, as the price of equity falls, the reduced equity must reflect the same volatility of the firm's cash flows and so negative returns should lead to increases in equity volatility. The leverage effect is pervasive in equity returns, especially in broad equity indices, although alone it is insufficient to explain the time variation of volatility (Christie, 1982; Bekaert and Wu, 2000).
- *Volatility Feedback*: Volatility feedback is motivated by a model where the volatility of an asset is priced. When the price of an asset falls, the volatility must increase to reflect the increased expected return (in the future) of this asset, and an increase in volatility requires an even lower price to generate a sufficient return to compensate an investor for holding a volatile asset. There is evidence that empirically supports this explantion although this feature alone cannot explain the totality of the time-variation of volatility (Bekaert and Wu, 2000).
- *Illiquidity*: Short run spells of illiquidity may produce time-varying volatility even when shocks are i.i.d. Intuitively, if the market is oversold (bought), a small negative (positive) shock produces a small decrease (increase) in demand. However, since few participants are willing to buy (sell), this shock has a disproportionate effect on prices. Liquidity runs tend to last from 20 minutes to a few days and cannot explain the long cycles in present volatility.
- *State Uncertainty*: Asset prices are essential instruments that allow agents to express beliefs about the state of the economy. When the state is uncertain, slight changes in beliefs may cause significant shifts in portfolio holdings which in turn feedback into beliefs about the state. This feedback loop can generate time-varying volatility and should have the most substantial effect when the economy is transitioning between periods of growth and contraction (Veronesi, 1999; Collard et al., 2018).

The economic causes of the time-variation in volatility include all of these and some not yet identified, such as behavioral causes.

7.1.1 What is volatility?

Volatility comes in many shapes and forms. It is critical to distinguish between related but different uses of “volatility”.

Volatility Volatility is the standard deviation. Volatility is often preferred to variance as it is measured in the same *units* as the original data. For example, when using returns, the volatility is also measured in returns, and so volatility of 5% indicates that $\pm 5\%$ is a meaningful quantity.

Realized Volatility Realized volatility has historically been used to denote a measure of the volatility over some arbitrary period of time,

$$\hat{\sigma} = \sqrt{T^{-1} \sum_{t=1}^T (r_t - \hat{\mu})^2} \quad (7.1)$$

but is now used to describe a volatility measure constructed using ultra-high-frequency (UHF) data (also known as tick data). See section 7.8 for details.

Conditional Volatility Conditional volatility is the expected volatility at some future time $t + h$ based on all available information up to time t (\mathcal{F}_t). The one-period ahead conditional volatility is denoted $E_t[\sigma_{t+1}]$.

Implied Volatility Implied volatility is the volatility that correctly prices an option. The Black-Scholes pricing formula relates the price of a European call option to the current price of the underlying, the strike, the risk-free rate, the time-to-maturity, and the *volatility*,

$$BS(S_t, K, r, t, \sigma_t) = C_t$$

where C is the price of the call. The implied volatility is the value which solves the Black-Scholes taking the option and underlying prices, the strike, the risk-free and the time-to-maturity as given,

$$\hat{\sigma}_t(S_t, K, r, t, C).$$

Recent econometric developments have produced nonparametric estimators that do not make strong assumptions on the underlying price process. The VIX is a leading example of these Model-free Implied Volatility (MFIV) estimators.

Annualized Volatility When volatility is measured over an interval other than a year, such as a day, week or month, it can always be scaled to reflect the volatility of the asset over a year. For example, if σ denotes the daily volatility of an asset and there are 252 trading days in a year, the annualized volatility is $\sqrt{252}\sigma$. Annualized volatility is a useful measure that removes the sampling interval from reported volatilities.

Variance All of the uses of volatility can be replaced with variance, and this chapter focuses on modeling *conditional variance* denoted $E_t[\sigma_{t+1}^2]$, or in the general case, $E_t[\sigma_{t+h}^2]$.

7.2 ARCH Models

In financial econometrics, an arch is not an architectural feature of a building; it is a fundamental tool for analyzing the time-variation of conditional variance. The success of the *ARCH* (AutoRegressive Conditional Heteroskedasticity) family of models can be attributed to three features: ARCH processes are essentially ARMA models, and many of the tools of linear time-series analysis can be directly applied, ARCH-family models are easy to estimate, and simple, parsimonious models are capable of accurate descriptions of the dynamics of asset volatility.

7.2.1 The ARCH model

The complete ARCH(P) model (Engle, 1982) relates the current level of volatility to the past P squared shocks.

Definition 7.1 (P^{th} Order Autoregressive Conditional Heteroskedasticity (ARCH)). A P^{th} order ARCH process is given by

$$\begin{aligned} r_t &= \mu_t + \varepsilon_t \\ \sigma_t^2 &= \omega + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \dots + \alpha_P \varepsilon_{t-P}^2 \\ \varepsilon_t &= \sigma_t e_t \\ e_t &\stackrel{\text{i.i.d.}}{\sim} N(0, 1). \end{aligned} \tag{7.2}$$

where μ_t can be any adapted model for the conditional mean.¹

The key feature of this model is that the variance of the shock, ε_t , is time varying and depends on the past P shocks, $\varepsilon_{t-1}, \varepsilon_{t-2}, \dots, \varepsilon_{t-P}$, through their squares. σ_t^2 is the time $t - 1$ conditional variance. All of the right-hand side variables that determine σ_t^2 are known at time $t - 1$, and so σ_t^2 is in the time $t - 1$ information set \mathcal{F}_{t-1} . The model for the conditional mean can include own lags, shocks (in an MA model) or exogenous variables such as the default spread or term premium. In practice, the model for the conditional mean should be general enough to capture the dynamics present in the data. In many financial time series, particularly when returns are measured over short intervals - one day to one week - a constant mean, sometimes assumed to be 0, is sufficient.

An common alternative description an ARCH(P) model is

$$\begin{aligned} r_t | \mathcal{F}_{t-1} &\sim N(\mu_t, \sigma_t^2) \\ \sigma_t^2 &= \omega + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \dots + \alpha_P \varepsilon_{t-P}^2 \\ \varepsilon_t &= r_t - \mu_t \end{aligned} \tag{7.3}$$

which is read “ r_t given the information set at time $t - 1$ is conditionally normal with mean μ_t and variance σ_t^2 ”.²

The conditional variance, σ_t^2 , is

$$E_{t-1} [\varepsilon_t^2] = E_{t-1} [e_t^2 \sigma_t^2] = \sigma_t^2 E_{t-1} [e_t^2] = \sigma_t^2 \tag{7.4}$$

and the unconditional variance, $\bar{\sigma}^2$, is

$$E [\varepsilon_{t+1}^2] = \bar{\sigma}^2. \tag{7.5}$$

The first interesting property of the ARCH(P) model is the unconditional variance. Assuming the

¹A model is adapted if everything required to model the mean at time t is known at time $t - 1$. Standard examples of adapted mean processes include a constant mean, ARMA processes or models containing exogenous regressors known at time $t - 1$.

²It is implausible that the unconditional (long-run) mean return of many risky assets is zero. However, when using daily equity data, the squared mean is typically less than 1% of the variance ($\frac{\mu^2}{\sigma^2} < 0.01$) and there are few consequences for setting the conditional mean to 0. Some assets, e.g., electricity prices, have non-trivial predictability and an appropriate model for the conditional mean is required before modeling the volatility.

unconditional variance exists, $\bar{\sigma}^2 = E[\sigma_t^2]$ can be derived from

$$\begin{aligned} E[\sigma_t^2] &= E[\omega + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \dots + \alpha_P \varepsilon_{t-P}^2] \\ &= \omega + \alpha_1 E[\varepsilon_{t-1}^2] + \alpha_2 E[\varepsilon_{t-2}^2] + \dots + \alpha_P E[\varepsilon_{t-P}^2] \\ &= \omega + \alpha_1 E[E_{t-2}[\sigma_{t-1}^2 e_{t-1}^2]] + \alpha_2 E[E_{t-3}[\sigma_{t-2}^2 e_{t-2}^2]] \\ &\quad + \dots + \alpha_P E[E_{t-P-1}[\sigma_{t-P}^2 e_{t-P}^2]] \\ &= \omega + \alpha_1 E[\sigma_{t-1}^2 E_{t-2}[e_{t-1}^2]] + \alpha_2 E[\sigma_{t-2}^2 E_{t-3}[e_{t-2}^2]] \\ &\quad + \dots + \alpha_P E[\sigma_{t-P}^2 E_{t-P-1}[e_{t-P}^2]] \\ &= \omega + \alpha_1 E[\sigma_{t-1}^2 \times 1] + \alpha_2 E[\sigma_{t-2}^2 \times 1] + \dots + \alpha_P E[\sigma_{t-P}^2 \times 1] \\ &= \omega + \alpha_1 E[\sigma_{t-1}^2] + \alpha_2 E[\sigma_{t-2}^2] + \dots + \alpha_P E[\sigma_{t-P}^2] \\ &= \omega + \alpha_1 E[\sigma_t^2] + \alpha_2 E[\sigma_t^2] + \dots + \alpha_P E[\sigma_t^2] \end{aligned} \tag{7.6}$$

$$\begin{aligned} E[\sigma_t^2](1 - \alpha_1 - \alpha_2 - \dots - \alpha_P) &= \omega \\ \bar{\sigma}^2 &= \frac{\omega}{1 - \alpha_1 - \alpha_2 - \dots - \alpha_P}. \end{aligned} \tag{7.7}$$

This derivation makes use of a number of properties of ARCH family models. First, the definition of the shock $\varepsilon_t^2 \equiv e_t^2 \sigma_t^2$ is used to separate the i.i.d. normal innovation (e_t) from the conditional variance (σ_t^2) using the Law of Iterated Expectations. For example, σ_{t-1} is known at time $t-2$ and so is in \mathcal{F}_{t-2} and $E_{t-1}[\sigma_{t-1}] = \sigma_{t-1}$. e_{t-1} is an i.i.d. draw at time $t-1$, a random variable at time $t-2$, and so $E_{t-2}[e_{t-1}^2] = 1$. The result follows from the property that the unconditional expectation of σ_{t-j}^2 is the same in any time period ($E[\sigma_t^2] = E[\sigma_{t-p}^2] = \bar{\sigma}^2$) in a covariance stationary time series. Inspection of the final line in the derivation reveals the condition needed to ensure that the unconditional expectation is finite: $1 - \alpha_1 - \alpha_2 - \dots - \alpha_P > 0$. As was the case in an AR model, as the persistence (as measured by $\alpha_1, \alpha_2, \dots$) increases towards a unit root, the process explodes.

7.2.1.1 Stationarity

An ARCH(P) model is covariance stationary as long as the model for the conditional mean corresponds to a stationary process³ and $1 - \alpha_1 - \alpha_2 - \dots - \alpha_P > 0$.⁴ ARCH models have the property that $E[\varepsilon_t^2] = \bar{\sigma}^2 = \omega / (1 - \alpha_1 - \alpha_2 - \dots - \alpha_P)$ since

$$E[\varepsilon_t^2] = E[e_t^2 \sigma_t^2] = E[E_{t-1}[e_t^2 \sigma_t^2]] = E[\sigma_t^2 E_{t-1}[e_t^2]] = E[\sigma_t^2 \times 1] = E[\sigma_t^2]. \tag{7.8}$$

which exploits the conditional (on \mathcal{F}_{t-1}) independence of e_t from σ_t^2 and the assumption that e_t is a mean zero process with unit variance so that $E[e_t^2] = 1$.

One crucial requirement of any covariance stationary ARCH process is that the parameters of the variance evolution, $\alpha_1, \alpha_2, \dots, \alpha_P$ must all be positive.⁵ The intuition behind this requirement is that if one of the α s were negative, eventually a shock would be sufficiently large to produce a negative

³For example, a constant or a covariance stationary ARMA process.

⁴When $\sum_{i=1}^P \alpha_i > 1$, and ARCH(P) may still be strictly stationary although it cannot be covariance stationary since it has infinite variance.

⁵Since each $\alpha_j \geq 0$, the roots of the characteristic polynomial associated with $\alpha_1, \alpha_2, \dots, \alpha_p$ are less than 1 if and only if $\sum_{p=1}^P \alpha_p < 1$.

conditional variance and an ill-defined process. Finally, it is also necessary that $\omega > 0$ to ensure covariance stationarity.

To aid in developing intuition about ARCH-family models consider a simple ARCH(1) with a constant mean of 0,

$$\begin{aligned} r_t &= \varepsilon_t \\ \sigma_t^2 &= \omega + \alpha_1 \varepsilon_{t-1}^2 \\ \varepsilon_t &= \sigma_t e_t \\ e_t &\stackrel{\text{i.i.d.}}{\sim} N(0, 1). \end{aligned} \tag{7.9}$$

While the conditional variance of an ARCH process appears different from anything previously encountered, the squared error ε_t^2 can be equivalently expressed as an AR(1). This transformation allows many properties of ARCH residuals to be directly derived by applying the results of chapter 4. By adding $\varepsilon_t^2 - \sigma_t^2$ to both sides of the volatility equation,

$$\begin{aligned} \sigma_t^2 &= \omega + \alpha_1 \varepsilon_{t-1}^2 \\ \sigma_t^2 + \varepsilon_t^2 - \sigma_t^2 &= \omega + \alpha_1 \varepsilon_{t-1}^2 + \varepsilon_t^2 - \sigma_t^2 \\ \varepsilon_t^2 &= \omega + \alpha_1 \varepsilon_{t-1}^2 + \varepsilon_t^2 - \sigma_t^2 \\ \varepsilon_t^2 &= \omega + \alpha_1 \varepsilon_{t-1}^2 + \sigma_t^2 (e_t^2 - 1) \\ \varepsilon_t^2 &= \omega + \alpha_1 \varepsilon_{t-1}^2 + v_t, \end{aligned} \tag{7.10}$$

an ARCH(1) process can be shown to be an AR(1). The error term, v_t represents the volatility *surprise*, $\varepsilon_t^2 - \sigma_t^2$, which can be decomposed as $\sigma_t^2(e_t^2 - 1)$. The shock is a mean 0 white noise process since e_t is i.i.d. and $E[e_t^2] = 1$. Using the AR representation, the autocovariances of ε_t^2 are simple to derive. First note that $\varepsilon_t^2 - \bar{\sigma}^2 = \sum_{i=0}^{\infty} \alpha_1^i v_{t-i}$. The first autocovariance can be expressed

$$\begin{aligned} E[(\varepsilon_t^2 - \bar{\sigma}^2)(\varepsilon_{t-1}^2 - \bar{\sigma}^2)] &= E \left[\left(\sum_{i=0}^{\infty} \alpha_1^i v_{t-i} \right) \left(\sum_{j=1}^{\infty} \alpha_1^{j-1} v_{t-j} \right) \right] \\ &= E \left[\left(v_t + \sum_{i=1}^{\infty} \alpha_1^i v_{t-i} \right) \left(\sum_{j=1}^{\infty} \alpha_1^{j-1} v_{t-j} \right) \right] \\ &= E \left[\left(v_t + \alpha_1 \sum_{i=1}^{\infty} \alpha_1^{i-1} v_{t-i} \right) \left(\sum_{j=1}^{\infty} \alpha_1^{j-1} v_{t-j} \right) \right] \\ &= E \left[v_t \left(\sum_{i=1}^{\infty} \alpha_1^{i-1} v_{t-i} \right) \right] + E \left[\alpha_1 \left(\sum_{i=1}^{\infty} \alpha_1^{i-1} v_{t-i} \right) \left(\sum_{j=1}^{\infty} \alpha_1^{j-1} v_{t-j} \right) \right] \\ &= \sum_{i=1}^{\infty} \alpha_1^{i-1} E[v_t v_{t-i}] + E \left[\alpha_1 \left(\sum_{i=1}^{\infty} \alpha_1^{i-1} v_{t-i} \right) \left(\sum_{j=1}^{\infty} \alpha_1^{j-1} v_{t-j} \right) \right] \end{aligned} \tag{7.11}$$

$$\begin{aligned}
&= \sum_{i=1}^{\infty} \alpha_1^{i-1} \cdot 0 + E \left[\alpha_1 \left(\sum_{i=1}^{\infty} \alpha_1^{i-1} v_{t-i} \right) \left(\sum_{j=1}^{\infty} \alpha_1^{j-1} v_{t-j} \right) \right] \\
&= \alpha_1 E \left[\left(\sum_{i=1}^{\infty} \alpha_1^{i-1} v_{t-i} \right)^2 \right] \\
&= \alpha_1 E \left[\left(\sum_{i=0}^{\infty} \alpha_1^i v_{t-1-i} \right)^2 \right] \\
&= \alpha_1 \left(\sum_{i=0}^{\infty} \alpha_1^{2i} E[v_{t-1-i}^2] + 2 \sum_{j=0}^{\infty} \sum_{k=j+1}^{\infty} \alpha_1^{jk} E[v_{t-1-j} v_{t-1-k}] \right) \\
&= \alpha_1 \left(\sum_{i=0}^{\infty} \alpha_1^{2i} V[v_{t-1-i}] + 2 \sum_{j=0}^{\infty} \sum_{k=j+1}^{\infty} \alpha_1^{jk} \cdot 0 \right) \\
&= \alpha_1 \sum_{i=0}^{\infty} \alpha_1^{2i} V[v_{t-1-i}] \\
&= \alpha_1 V[\varepsilon_{t-1}^2]
\end{aligned}$$

where $V[\varepsilon_{t-1}^2] = V[\varepsilon_t^2]$ is the variance of the squared innovations.⁶ By repeated substitution, the s^{th} autocovariance, $E[(\varepsilon_t^2 - \bar{\sigma}^2)(\varepsilon_{t-s}^2 - \bar{\sigma}^2)]$, can be shown to be $\alpha_1^s V[\varepsilon_t^2]$, and so that the autocovariances of an ARCH(1) process are identical to those of an AR(1) process.

7.2.1.2 Autocorrelations

Using the autocovariances, the autocorrelations are

$$\text{Corr}(\varepsilon_t^2, \varepsilon_{t-s}^2) = \frac{\alpha_1^s V[\varepsilon_t^2]}{V[\varepsilon_t^2]} = \alpha_1^s. \quad (7.12)$$

Further, the relationship between the s^{th} autocorrelation of an ARCH process and an AR process holds for ARCH processes with other orders. The autocorrelations of an ARCH(P) are identical to those of an AR(P) process with $\{\phi_1, \phi_2, \dots, \phi_P\} = \{\alpha_1, \alpha_2, \dots, \alpha_P\}$. One interesting aspect of ARCH(P) processes (and any covariance stationary ARCH-family model) is that the autocorrelations of $\{\varepsilon_t^2\}$ must be positive. If one autocorrelation were negative, eventually a shock would be sufficiently large to force the conditional variance negative, and so the process would be ill-defined. In practice it is often better to examine the absolute values ($\text{Corr}(|\varepsilon_t|, |\varepsilon_{t-s}|)$) rather than the squares since financial returns frequently have outliers that are exacerbated when squared.

7.2.1.3 Kurtosis

The second interesting property of ARCH models is that the kurtosis of shocks (ε_t) is strictly greater than the kurtosis of a normal. This may seem strange since all of the shocks $\varepsilon_t = \sigma_t e_t$ are normal by

⁶For the time being, assume this is finite.

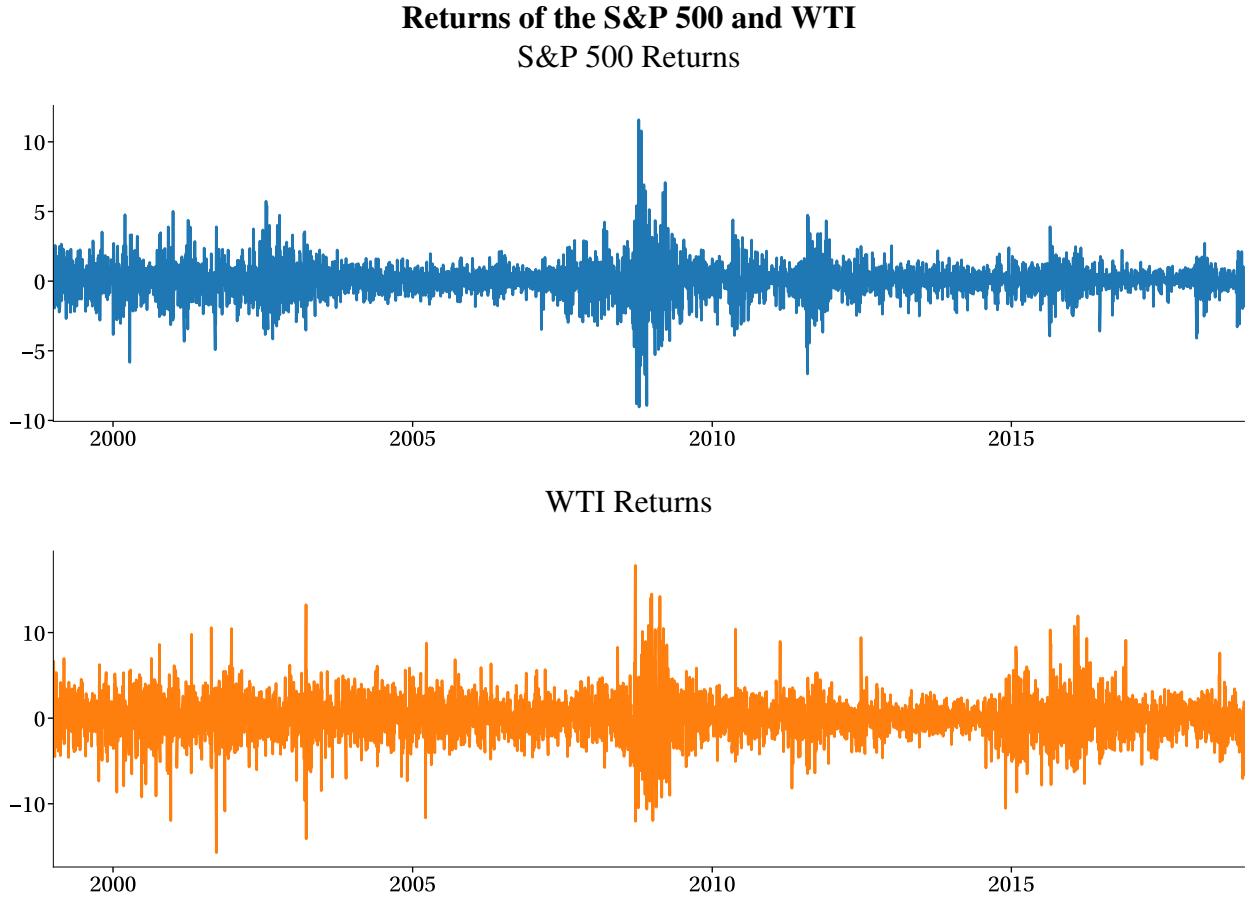


Figure 7.1: Plots of S&P 500 and WTI returns (scaled by 100) from 1999 until 2018. The bulges in the return plots are graphical evidence of time-varying volatility.

assumption. An ARCH model is a *variance-mixture* of normals, and so must have a kurtosis larger than three. The direct proof is simple,

$$\kappa = \frac{E[\varepsilon_t^4]}{E[\varepsilon_t^2]^2} = \frac{E[E_{t-1}[\varepsilon_t^4]]}{E[E_{t-1}[e_t^2 \sigma_t^2]]^2} = \frac{E[E_{t-1}[e_t^4] \sigma_t^4]}{E[E_{t-1}[e_t^2] \sigma_t^2]^2} = \frac{E[3\sigma_t^4]}{E[\sigma_t^2]^2} = 3 \frac{E[\sigma_t^4]}{E[\sigma_t^2]^2} \geq 3. \quad (7.13)$$

The key steps in this derivation are that $\varepsilon_t^4 = e_t^4 \sigma_t^4$ and that $E_t[e_t^4] = 3$ since e_t is a standard normal. The final conclusion that $E[\sigma_t^4]/E[\sigma_t^2]^2 > 1$ follows from noting that for any random variable Y , $V[Y] = E[Y^2] - E[Y]^2 \geq 0$ and so it must be the case that $E[\sigma_t^4] \geq E[\sigma_t^2]^2$ or $\frac{E[\sigma_t^4]}{E[\sigma_t^2]^2} \geq 1$. The kurtosis, κ , of an ARCH(1) can be shown to be

$$\kappa = \frac{3(1 - \alpha_1^2)}{(1 - 3\alpha_1^2)} > 3 \quad (7.14)$$

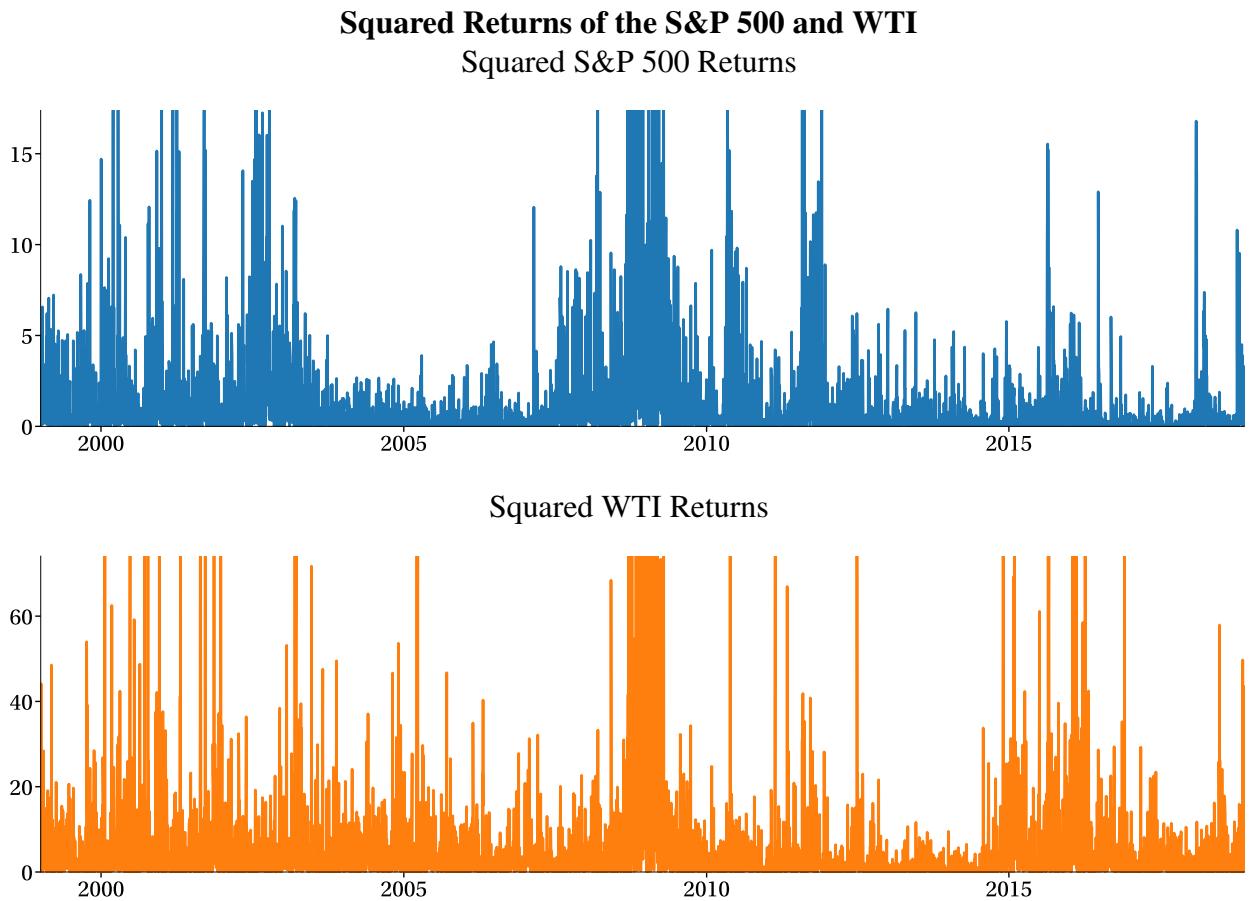


Figure 7.2: Plots of the squared returns of the S&P 500 Index and WTI. Time-variation in the squared returns is evidence of ARCH.

which is greater than 3 since $1 - 3\alpha_1^2 < 1 - \alpha_1^2$ for any value of $\alpha \neq 0$. The complete derivation of the kurtosis is involved and is presented in Appendix 7.A.

7.2.2 The GARCH model

The ARCH model has been deemed a sufficient contribution to economics to warrant a Nobel prize. Unfortunately, like most models, it has problems. ARCH models typically require 5-8 lags of the squared shock to model conditional variance adequately. The Generalized ARCH (GARCH) process, introduced by Bollerslev (1986), improves the original specification adding lagged conditional variance, which acts as a *smoothing* term. A low-order GARCH model typically fits as well as a high-order ARCH.

Definition 7.2 (Generalized Autoregressive Conditional Heteroskedasticity (GARCH) process). A GARCH(P,Q) process is defined as

$$r_t = \mu_t + \varepsilon_t \quad (7.15)$$

$$\begin{aligned}\sigma_t^2 &= \omega + \sum_{p=1}^P \alpha_p \varepsilon_{t-p}^2 + \sum_{q=1}^Q \beta_q \sigma_{t-q}^2 \\ \varepsilon_t &= \sigma_t e_t \\ e_t &\stackrel{\text{i.i.d.}}{\sim} N(0, 1)\end{aligned}$$

where μ_t can be any adapted model for the conditional mean.

The GARCH(P,Q) model builds on the ARCH(P) model by including Q lags of the conditional variance, $\sigma_{t-1}^2, \sigma_{t-2}^2, \dots, \sigma_{t-Q}^2$. Rather than focusing on the general specification with all of its complications, consider a simpler GARCH(1,1) model where the conditional mean is assumed to be zero,

$$\begin{aligned}r_t &= \varepsilon_t \\ \sigma_t^2 &= \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \\ \varepsilon_t &= \sigma_t e_t \\ e_t &\stackrel{\text{i.i.d.}}{\sim} N(0, 1)\end{aligned}\tag{7.16}$$

In this specification, the future variance will be an average of the current shock, ε_{t-1}^2 , the current variance, σ_{t-1}^2 , and a constant. Including the lagged variance produces a model that can be equivalently expressed as an ARCH(∞). Begin by backward substituting for σ_{t-1}^2 ,

$$\begin{aligned}\sigma_t^2 &= \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \\ &= \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta \underbrace{(\omega + \alpha_1 \varepsilon_{t-2}^2 + \beta_1 \sigma_{t-2}^2)}_{\sigma_{t-1}^2} \\ &= \omega + \beta_1 \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \alpha_1 \varepsilon_{t-2}^2 + \beta_1^2 \sigma_{t-2}^2 \\ &= \omega + \beta_1 \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \alpha_1 \varepsilon_{t-2}^2 + \beta_1^2 \underbrace{\omega + \alpha_1 \varepsilon_{t-3}^2 + \beta_1 \sigma_{t-3}^2}_{\sigma_{t-2}^2} \\ &= \omega + \beta_1 \omega + \beta_1^2 \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \alpha_1 \varepsilon_{t-2}^2 + \beta_1^2 \alpha_1 \varepsilon_{t-3}^2 + \beta_1^3 \sigma_{t-3}^2 \\ &= \sum_{i=0}^{\infty} \beta_1^i \omega + \sum_{i=0}^{\infty} \beta_1^i \alpha_1 \varepsilon_{t-i-1}^2,\end{aligned}\tag{7.17}$$

and the ARCH(∞) representation can be derived.⁷ The conditional variance in period t depends on a constant, $\sum_{i=0}^{\infty} \beta_1^i \omega = \frac{\omega}{1-\beta}$, and a weighted average of past squared innovations with weights $\alpha_1, \beta_1 \alpha_1, \beta_1^2 \alpha_1, \beta_1^3 \alpha_1, \dots$.

As was the case in the ARCH(P) model, the coefficients of a GARCH model must be restricted to ensure the conditional variances are uniformly positive. In a GARCH(1,1), these restrictions are $\omega > 0, \alpha_1 \geq 0$ and $\beta_1 \geq 0$. In a GARCH(P,1) model the restriction change to $\alpha_p \geq 0, p = 1, 2, \dots, P$ with the same restrictions on ω and β_1 . The minimal parameter restrictions needed to ensure that

⁷Since the model is assumed to be stationary, it must be the case that $0 \leq \beta < 1$ and so $\lim_{j \rightarrow \infty} \beta^j \sigma_{t-j}^2 = 0$.

variances are always positive are difficult to derive for the full class of GARCH(P,Q) models. For example, in a GARCH(2,2), one of the two β 's (β_2) can be slightly negative while ensuring that all conditional variances are positive. See Nelson and Cao (1992) for further details.

The GARCH(1,1) model can be transformed into a standard time series model for ε_t^2 by adding $\varepsilon_t^2 - \sigma_t^2$ to both sides.

$$\begin{aligned}\sigma_t^2 &= \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \\ \sigma_t^2 + \varepsilon_t^2 - \sigma_t^2 &= \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 + \varepsilon_t^2 - \sigma_t^2 \\ \varepsilon_t^2 &= \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 + \varepsilon_t^2 - \sigma_t^2 \\ \varepsilon_t^2 &= \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 - \beta_1 \varepsilon_{t-1}^2 + \beta_1 \varepsilon_{t-1}^2 + \varepsilon_t^2 - \sigma_t^2 \\ \varepsilon_t^2 &= \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \varepsilon_{t-1}^2 - \beta_1 (\varepsilon_{t-1}^2 - \sigma_{t-1}^2) + \varepsilon_t^2 - \sigma_t^2 \\ \varepsilon_t^2 &= \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \varepsilon_{t-1}^2 - \beta_1 v_{t-1} + v_t \\ \varepsilon_t^2 &= \omega + (\alpha_1 + \beta_1) \varepsilon_{t-1}^2 - \beta_1 v_{t-1} + v_t\end{aligned}\tag{7.18}$$

The squared shock in a GARCH(1,1) follows an ARMA(1,1) process where $v_t = \varepsilon_t^2 - \sigma_t^2$ is the volatility surprise. In the general GARCH(P,Q), the ARMA representation takes the form of an ARMA(max(P,Q),Q).

$$\varepsilon_t^2 = \omega + \sum_{i=1}^{\max(P,Q)} (\alpha_i + \beta_i) \varepsilon_{t-i}^2 - \sum_{q=1}^Q \beta_1 v_{t-q} + v_t\tag{7.19}$$

The unconditional variance is computed by taking expectations of both sides, so that

$$\begin{aligned}E[\sigma_t^2] &= \omega + \alpha_1 E[\varepsilon_{t-1}^2] + \beta_1 E[\sigma_{t-1}^2] \\ \bar{\sigma}^2 &= \omega + \alpha_1 \bar{\sigma}^2 + \beta_1 \bar{\sigma}^2 \\ \bar{\sigma}^2 - \alpha_1 \bar{\sigma}^2 - \beta_1 \bar{\sigma}^2 &= \omega \\ \bar{\sigma}^2 &= \frac{\omega}{1 - \alpha_1 - \beta_1}.\end{aligned}\tag{7.20}$$

Inspection of the ARMA model leads to an alternative derivation of $\bar{\sigma}^2$ since the AR coefficient is $\alpha_1 + \beta_1$ and the intercept is ω , and the unconditional mean in an ARMA(1,1) is the intercept divided by one minus the AR coefficient, $\omega/(1 - \alpha_1 - \beta_1)$. In a general GARCH(P,Q) the unconditional variance is

$$\bar{\sigma}^2 = \frac{\omega}{1 - \sum_{p=1}^P \alpha_p - \sum_{q=1}^Q \beta_q}.\tag{7.21}$$

The requirements on the parameters for stationarity in a GARCH(1,1) are $1 - \alpha_1 - \beta > 0$ and $\alpha_1 \geq 0$, $\beta_1 \geq 0$ and $\omega > 0$.

The ARMA(1,1) form can be used directly to solve for the autocovariances. Recall the definition of a mean zero ARMA(1,1),

$$Y_t = \phi Y_{t-1} + \theta \varepsilon_{t-1} + \varepsilon_t\tag{7.22}$$

The 1st autocovariance can be computed as

$$\begin{aligned} E[Y_t Y_{t-1}] &= E[(\phi Y_{t-1} + \theta \varepsilon_{t-1} + \varepsilon_t) Y_{t-1}] \\ &= E[\phi Y_{t-1}^2] + [\theta \varepsilon_{t-1}^2] \\ &= \phi V[Y_{t-1}] + \theta V[\varepsilon_{t-1}] \\ \gamma_1 &= \phi V[Y_{t-1}] + \theta V[\varepsilon_{t-1}] \end{aligned} \quad (7.23)$$

and the s^{th} autocovariance is $\gamma_s = \phi^{s-1} \gamma_1$. In the notation of a GARCH(1,1) model, $\phi = \alpha_1 + \beta_1$, $\theta = -\beta_1$, Y_{t-1} is ε_{t-1}^2 and η_{t-1} is $\sigma_{t-1}^2 - \varepsilon_{t-1}^2$. Thus, $V[\varepsilon_{t-1}^2]$ and $V[\sigma_t^2 - \varepsilon_t^2]$ must be solved for. This derivation is challenging and so is presented in the appendix. The key to understanding the autocovariance (and autocorrelation) of a GARCH is to use the ARMA mapping. First note that $E[\sigma_t^2 - \varepsilon_t^2] = 0$ so $V[\sigma_t^2 - \varepsilon_t^2]$ is simply $E[(\sigma_t^2 - \varepsilon_t^2)^2]$. This can be expanded to $E[\varepsilon_t^4] - 2E[\varepsilon_t^2 \sigma_t^2] + E[\sigma_t^4]$ which can be shown to be $2E[\sigma_t^4]$. The only remaining step is to complete the tedious derivation of the expectation of these fourth powers which is presented in Appendix 7.B.

7.2.2.1 Kurtosis

The kurtosis can be shown to be

$$\kappa = \frac{3(1 + \alpha_1 + \beta_1)(1 - \alpha_1 - \beta_1)}{1 - 2\alpha_1\beta_1 - 3\alpha_1^2 - \beta_1^2} > 3. \quad (7.24)$$

The kurtosis is larger than that of a normal despite the innovations, e_t , all having normal distributions since that model is a variance mixture of normals. The formal derivation is presented in 7.B.

Exponentially Weighted Moving Averages (EWMA)

Exponentially Weighted Moving Averages, popularized by RiskMetrics, are commonly used to measure and forecast volatilities from returns without estimating any parameters (J.P.Morgan/Reuters, 1996). An EWMA is a restricted GARCH(1,1) model where $\omega = 0$ and $\alpha + \beta = 1$. The recursive form of an EWMA is

$$\sigma_t^2 = (1 - \lambda) \varepsilon_{t-1}^2 + \lambda \sigma_{t-1}^2,$$

which can be equivalently expressed as an ARCH(∞)

$$\sigma_t^2 = (1 - \lambda) \sum_{i=0}^{\infty} \lambda^i \varepsilon_{t-i-1}^2.$$

The weights on the lagged squared returns decay exponentially so that the ratio of two consecutive weights is λ . The single parameter λ is typically set to 0.94 when using daily returns, 0.97 when using weekly return data, or 0.99 when using monthly returns. These values were calibrated on a wide range of assets to forecast volatility well.

7.2.3 The EGARCH model

The Exponential GARCH (EGARCH) model represents a major shift from the ARCH and GARCH models (Nelson, 1991). Rather than model the variance directly, EGARCH models the natural log-

arithm of the variance, and so no parameters restrictions are required to ensure that the conditional variance is positive.

Definition 7.3 (Exponential Generalized Autoregressive Conditional Heteroskedasticity (EGARCH) process). An EGARCH(P,O,Q) process is defined

$$\begin{aligned} r_t &= \mu_t + \varepsilon_t \\ \ln(\sigma_t^2) &= \omega + \sum_{p=1}^P \alpha_p \left(\left| \frac{\varepsilon_{t-p}}{\sigma_{t-p}} \right| - \sqrt{\frac{2}{\pi}} \right) + \sum_{o=1}^O \gamma_o \frac{\varepsilon_{t-o}}{\sigma_{t-o}} + \sum_{q=1}^Q \beta_q \ln(\sigma_{t-q}^2) \\ \varepsilon_t &= \sigma_t e_t \\ e_t &\stackrel{\text{i.i.d.}}{\sim} N(0, 1) \end{aligned} \tag{7.25}$$

where μ_t can be any adapted model for the conditional mean. P and O were assumed to be equal in the original parameterization of Nelson (1991).

Rather than working with the complete specification, consider a simpler version, an EGARCH(1,1,1) with a constant mean,

$$\begin{aligned} r_t &= \mu + \varepsilon_t \\ \ln(\sigma_t^2) &= \omega + \alpha_1 \left(\left| \frac{\varepsilon_{t-1}}{\sigma_{t-1}} \right| - \sqrt{\frac{2}{\pi}} \right) + \gamma_1 \frac{\varepsilon_{t-1}}{\sigma_{t-1}} + \beta_1 \ln(\sigma_{t-1}^2) \\ \varepsilon_t &= \sigma_t e_t \\ e_t &\stackrel{\text{i.i.d.}}{\sim} N(0, 1). \end{aligned} \tag{7.26}$$

Three terms drive the dynamics in the log variance. The first term, $\left| \frac{\varepsilon_{t-1}}{\sigma_{t-1}} \right| - \sqrt{\frac{2}{\pi}}$, is the absolute value of a normal random variable, e_{t-1} , minus its expectation, $\sqrt{2/\pi}$, and so it is a mean zero shock. The second term, e_{t-1} – a standard normal – is an additional mean zero shock and the final term is the lagged log variance. The two shocks behave differently: the absolute value in the first produces a symmetric rise in the log variance for a given return while the sign of the second produces an asymmetric effect. γ_1 is typically estimated to negative so that volatility rises more after negative shocks than after positive ones. In the usual case where $\gamma_1 < 0$, the magnitude of the shock can be decomposed by conditioning on the sign of e_{t-1}

$$\text{Shock coefficient} = \begin{cases} \alpha_1 + \gamma_1 & \text{when } e_{t-1} < 0 \\ \alpha_1 - \gamma_1 & \text{when } e_{t-1} > 0 \end{cases} \tag{7.27}$$

Since both shocks are mean zero and the current log variance is linearly related to past log variance through β_1 , the EGARCH(1,1,1) model is an AR model.

EGARCH models often provide superior fits when compared to standard GARCH models. The presence of the asymmetric term is largely responsible for the superior fit since many asset return series have been found to exhibit a “leverage” effect. Additionally, the use of standardized shocks (e_{t-1}) in the dynamics of the log-variance reduces the effect of outliers.

Summary Statistics		
	S&P 500	WTI
Ann. Mean	14.03	5.65
Ann. Volatility	38.57	19.04
Skewness	0.063	-0.028
Kurtosis	7.22	11.45

Table 7.1: Summary statistics for the S&P 500 and WTI. Means and volatilities are reported in annualized terms using $100 \times$ returns. Skewness and kurtosis are scale-free by definition.

7.2.3.1 The S&P 500 and West Texas Intermediate Crude

The application of GARCH models will be demonstrated using daily returns on both the S&P 500 and West Texas Intermediate (WTI) Crude spot prices from January 1, 1999, until December 31, 2018. The S&P 500 data is from Yahoo! Finance and the WTI data is from the St. Louis Federal Reserve's FRED database. All returns are scaled by 100. The returns are plotted in Figure 7.1, the squared returns are plotted in Figure 7.2, and the absolute values of the returns are plotted in Figure 7.3. The plots of the squared returns and the absolute values of the returns are useful graphical diagnostics for detecting ARCH. If the residuals are conditionally heteroskedastic, both plots provide evidence of volatility dynamics in the transformed returns. In practice, the plot of the absolute returns is a more helpful graphical tool than the plot of the squares. Squared returns are noisy proxies for the variance, and the dynamics in the data may be obscured by a small number of outliers.

Summary statistics are presented in table 7.1, and estimates from an ARCH(5), and GARCH(1,1) and an EGARCH(1,1,1) are presented in table 7.2. The summary statistics are typical of financial data where both series are heavy-tailed (leptokurtotic).

Definition 7.4 (Leptokurtosis). A random variable x_t is said to be leptokurtic if its kurtosis,

$$\kappa = \frac{E[(x_t - E[x_t])^4]}{E[(x_t - E[x_t])^2]^2}$$

is greater than that of a normal ($\kappa > 3$). Leptokurtic variables are also known as “heavy-tailed” or “fat tailed”.

Definition 7.5 (Platykurtosis). A random variable x_t is said to be platykurtic if its kurtosis,

$$\kappa = \frac{E[(x_t - E[x_t])^4]}{E[(x_t - E[x_t])^2]^2}$$

is less than that of a normal ($\kappa < 3$). Platykurtic variables are also known as “thin-tailed”.

Table 7.2 contains estimates from an ARCH(5), a GARCH(1,1) and an EGARCH(1,1,1) model. All estimates were computed using maximum likelihood assuming the innovations are conditionally normally distributed. There is strong evidence of time-varying variance since most p-values are near 0. The highest log-likelihood (a measure of fit) is produced by the EGARCH model in both series. This is likely due to the EGARCH's inclusion of asymmetries, a feature excluded from both the ARCH and GARCH models.

S&P 500						
ARCH(5)						
ω	α_1	α_2	α_3	α_4	α_5	Log Lik.
0.294 (0.000)	0.095 (0.000)	0.204 (0.000)	0.189 (0.000)	0.193 (0.000)	0.143 (0.000)	-7008
GARCH(1,1)						
ω	α_1	β_1				
0.018 (0.000)	0.102 (0.000)	0.885 (0.000)				
EGARCH(1,1,1)						
ω	α_1	γ_1	β_1			
0.000 (0.909)	0.136 (0.000)	-0.153 (0.000)	0.975 (0.000)			

WTI						
ARCH(5)						
ω	α_1	α_2	α_3	α_4	α_5	Log Lik.
2.282 (0.000)	0.138 (0.000)	0.129 (0.000)	0.131 (0.000)	0.094 (0.000)	0.130 (0.000)	-11129
GARCH(1,1)						
ω	α_1	β_1				
0.047 (0.034)	0.059 (0.000)	0.934 (0.000)				
EGARCH(1,1,1)						
ω	α_1	γ_1	β_1			
0.020 (0.002)	0.109 (0.000)	-0.050 (0.000)	0.990 (0.000)			

Table 7.2: Parameter estimates, p-values and log-likelihoods from ARCH(5), GARCH(1,1) and EGARCH(1,1,1) models for the S&P 500 and WTI. These parameter values are typical of models estimated on daily data. The persistence of conditional variance, as measured by the sum of the α s in the ARCH(5), $\alpha_1 + \beta_1$ in the GARCH(1,1) and β_1 in the EGARCH(1,1,1), is high in all models. The log-likelihoods indicate the EGARCH model is preferred for both return series.

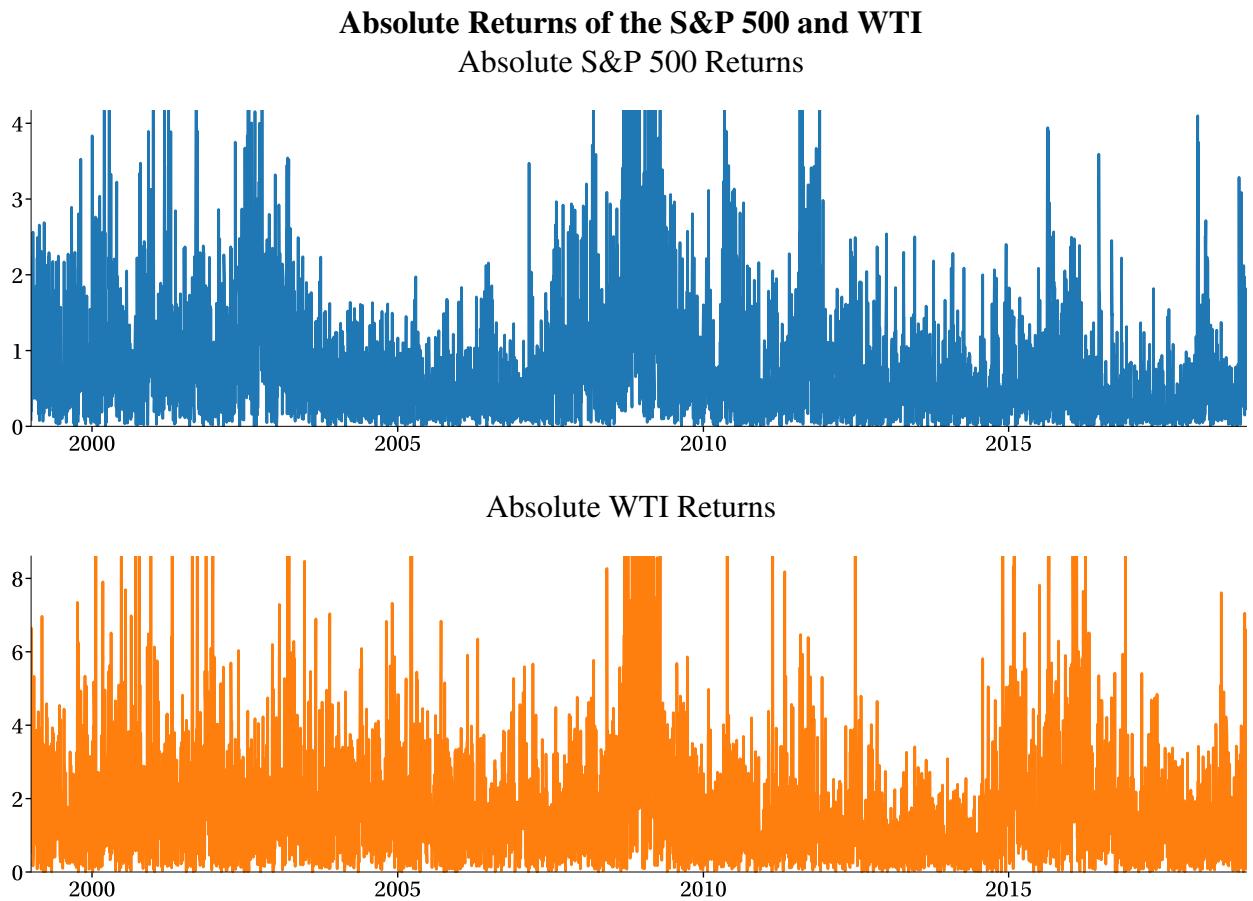


Figure 7.3: Plots of the absolute returns of the S&P 500 and WTI. Plots of the absolute value are often more useful in detecting ARCH as they are less noisy than squared returns yet still show changes in conditional volatility.

7.2.4 Alternative Specifications

Many extensions to the basic ARCH model have been introduced to capture important empirical features. This section outlines three of the most useful extensions in the ARCH-family.

7.2.4.1 GJR-GARCH

The GJR-GARCH model was named after the authors who introduced it, Glosten, Jagannathan, and Runkle (1993). It extends the standard GARCH(P,Q) by adding asymmetric terms that capture a common phenomenon in the conditional variance of equities: the propensity of the volatility to rise more after large negative shocks than to large positive shocks (known as the “leverage effect”).

Definition 7.6 (GJR-Generalized Autoregressive Conditional Heteroskedasticity (GJR-GARCH) process). A GJR-GARCH(P,O,Q) process is defined as

$$r_t = \mu_t + \varepsilon_t \quad (7.28)$$

$$\begin{aligned}\sigma_t^2 &= \omega + \sum_{p=1}^P \alpha_p \varepsilon_{t-p}^2 + \sum_{o=1}^O \gamma_o \varepsilon_{t-o}^2 I_{[\varepsilon_{t-o}<0]} + \sum_{q=1}^Q \beta_q \sigma_{t-q}^2 \\ \varepsilon_t &= \sigma_t e_t \\ e_t &\stackrel{\text{i.i.d.}}{\sim} N(0, 1)\end{aligned}$$

where μ_t can be any adapted model for the conditional mean, and $I_{[\varepsilon_{t-o}<0]}$ is an indicator function that takes the value 1 if $\varepsilon_{t-o} < 0$ and 0 otherwise.

The parameters of the GJR-GARCH, like the standard GARCH model, must be restricted to ensure that the fit variances are always positive. This set is difficult to describe for all GJR-GARCH(P,O,Q) models although it is simple of a GJR-GARCH(1,1,1). The dynamics in a GJR-GARCH(1,1,1) evolve according to

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \gamma_1 \varepsilon_{t-1}^2 I_{[\varepsilon_{t-1}<0]} + \beta_1 \sigma_{t-1}^2. \quad (7.29)$$

and it must be the case that $\omega > 0$, $\alpha_1 \geq 0$, $\alpha_1 + \gamma_1 \geq 0$ and $\beta_1 \geq 0$. If the innovations are conditionally normal, a GJR-GARCH model will be covariance stationary as long as the parameter restriction are satisfied and $\alpha_1 + \frac{1}{2}\gamma_1 + \beta_1 < 1$.

7.2.4.2 AVGARCH/TARCH/ZARCH

The Threshold ARCH (TARCH) model (also known as AVGARCH and ZARCH) makes one fundamental change to the GJR-GARCH model (Taylor, 1986; Zakoian, 1994). The TARCH model parameterizes the *conditional standard deviation* as a function of the lagged absolute value of the shocks. It also captures asymmetries using an asymmetric term that is similar to the asymmetry in the GJR-GARCH model.

Definition 7.7 (Threshold Autoregressive Conditional Heteroskedasticity (TARCH) process). A TARCH(P, O, Q) process is defined as

$$\begin{aligned}r_t &= \mu_t + \varepsilon_t \\ \sigma_t &= \omega + \sum_{p=1}^P \alpha_p |\varepsilon_{t-p}| + \sum_{o=1}^O \gamma_o |\varepsilon_{t-o}| I_{[\varepsilon_{t-o}<0]} + \sum_{q=1}^Q \beta_q \sigma_{t-q} \\ \varepsilon_t &= \sigma_t e_t \\ e_t &\stackrel{\text{i.i.d.}}{\sim} N(0, 1)\end{aligned} \quad (7.30)$$

where μ_t can be any adapted model for the conditional mean. TARCH models are also known as ZARCH due to Zakoian (1994) or AVGARCH when no asymmetric terms are included ($O = 0$, Taylor (1986)).

Below is an example of a TARCH(1,1,1) model.

$$\sigma_t = \omega + \alpha_1 |\varepsilon_{t-1}| + \gamma_1 |\varepsilon_{t-1}| I_{[\varepsilon_{t-1}<0]} + \beta_1 \sigma_{t-1}, \quad \alpha_1 + \gamma_1 \geq 0 \quad (7.31)$$

where $I_{[\varepsilon_{t-1}<0]}$ is an indicator variable which takes the value 1 if $\varepsilon_{t-1} < 0$. Models of the conditional standard deviation often outperform models that parameterize the conditional variance. The absolute shocks are less responsive than the squared shocks.

7.2.4.3 APARCH

The third model extends the TARCH and GJR-GARCH models by directly parameterizing the non-linearity in the conditional variance. Where the GJR-GARCH model uses 2, and the TARCH model uses 1, the Asymmetric Power ARCH (APARCH) of Ding, Granger, and Engle (1993) parameterizes this value directly (using δ). This form provides greater flexibility in modeling the memory of volatility while remaining parsimonious.

Definition 7.8 (Asymmetric Power Autoregressive Conditional Heteroskedasticity (APARCH) process). An APARCH(P,O,Q) process is defined as

$$\begin{aligned} r_t &= \mu_t + \varepsilon_t & (7.32) \\ \sigma_t^\delta &= \omega + \sum_{j=1}^{\max(P,O)} \alpha_j (|\varepsilon_{t-j}| + \gamma_j \varepsilon_{t-j})^\delta + \sum_{q=1}^Q \beta_q \sigma_{t-q}^\delta \\ \varepsilon_t &= \sigma_t e_t \\ e_t &\stackrel{\text{i.i.d.}}{\sim} N(0, 1) \end{aligned}$$

where μ_t can be any adapted model for the conditional mean. It must be the case that $P \geq O$ in an APARCH model, and if $P > O$, then $\gamma_j = 0$ for $j > O$. If that $\omega > 0$, $\alpha_k \geq 0$ and $-1 \leq \gamma_j \leq 1$, then the conditional variance are always positive.

It is not completely obvious to see that the APARCH model nests the GJR-GARCH and TARCH models as special cases. To examine how an APARCH nests a GJR-GARCH, consider an APARCH(1,1,1) model.

$$\sigma_t^\delta = \omega + \alpha_1 (|\varepsilon_{t-1}| + \gamma_1 \varepsilon_{t-1})^\delta + \beta_1 \sigma_{t-1}^\delta \quad (7.33)$$

Suppose $\delta = 2$, then

$$\begin{aligned} \sigma_t^2 &= \omega + \alpha_1 (|\varepsilon_{t-1}| + \gamma_1 \varepsilon_{t-1})^2 + \beta_1 \sigma_{t-1}^2 & (7.34) \\ &= \omega + \alpha_1 |\varepsilon_{t-1}|^2 + 2\alpha_1 \gamma_1 \varepsilon_{t-1} |\varepsilon_{t-1}| + \alpha_1 \gamma_1^2 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \\ &= \omega + \alpha_1 \varepsilon_{t-1}^2 + \alpha_1 \gamma_1^2 \varepsilon_{t-1}^2 + 2\alpha_1 \gamma_1 \varepsilon_{t-1}^2 \text{sign}(\varepsilon_{t-1}) + \beta_1 \sigma_{t-1}^2 \end{aligned}$$

where $\text{sign}(\cdot)$ is a function that returns 1 if its argument is positive and -1 if its argument is negative. Consider the total effect of ε_{t-1}^2 as it depends on the sign of ε_{t-1} ,

$$\text{Shock coefficient} = \begin{cases} \alpha_1 + \alpha_1 \gamma_1^2 + 2\alpha_1 \gamma_1 & \text{when } \varepsilon_{t-1} > 0 \\ \alpha_1 + \alpha_1 \gamma_1^2 - 2\alpha_1 \gamma_1 & \text{when } \varepsilon_{t-1} < 0 \end{cases} \quad (7.35)$$

γ is usually estimated to be less than zero which corresponds to the typical “leverage effect” in GJR-GARCH models.⁸ The relationship between a TARCH model and an APARCH model works analogously by setting $\delta = 1$. The APARCH model also nests the ARCH(P), GARCH(P,Q) and AV-GARCH(P,Q) models as special cases when $\gamma_1 = 0$.

⁸The explicit relationship between an APARCH and a GJR-GARCH can be derived when $\delta = 2$ by solving a system of two equation in two unknowns where eq. (7.35) is equated with the effect in a GJR-GARCH model.

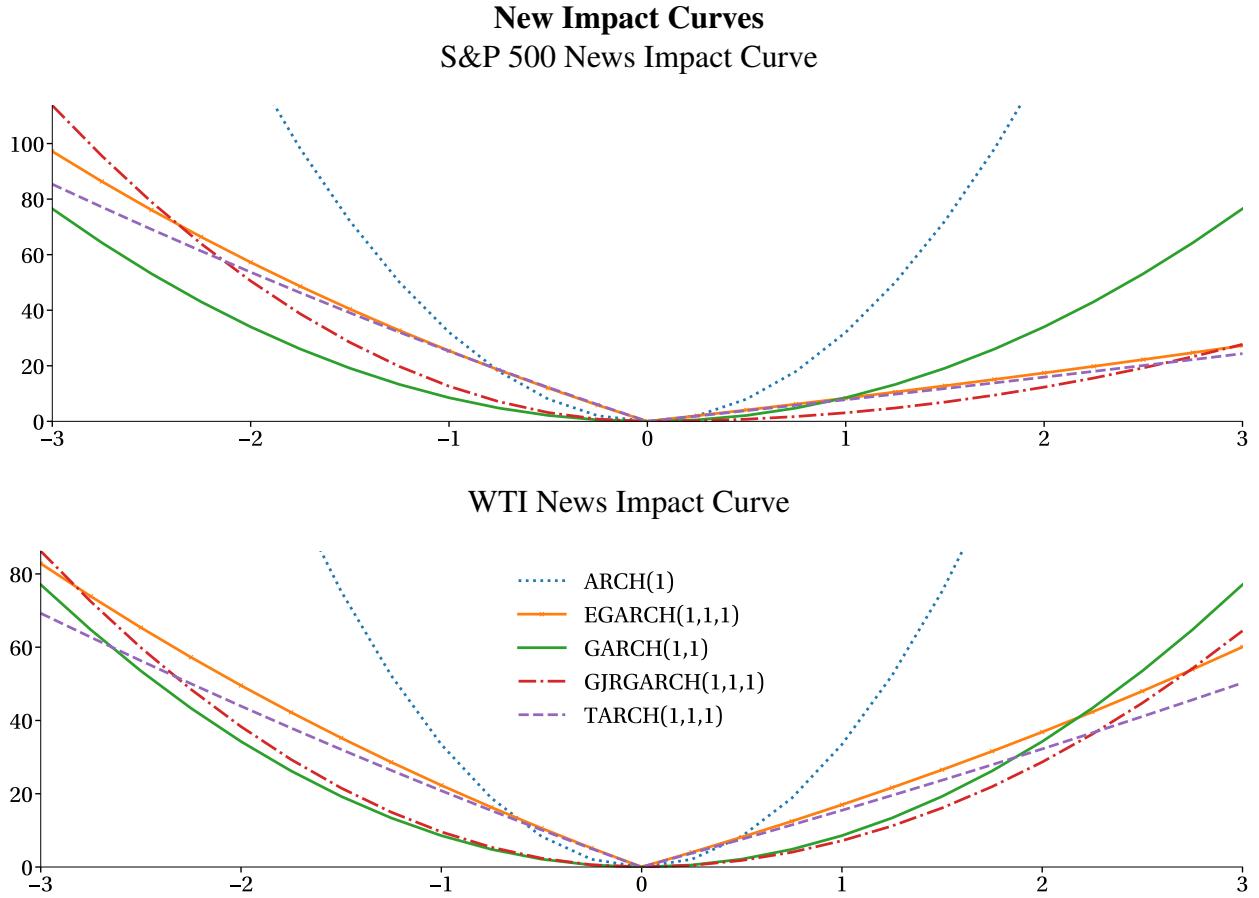


Figure 7.4: News impact curves for returns on both the S&P 500 and WTI. While the ARCH and GARCH curves are symmetric, the others show substantial asymmetries to negative news. Additionally, the fit APARCH models chose $\hat{\delta} \approx 1$, and so the NIC of the APARCH and the TARCH models appear similar.

7.2.5 The News Impact Curve

With a wide range of volatility models, each with a different specification for the dynamics of conditional variances, it can be difficult to determine the precise effect of a shock to the conditional variance. *News impact curves* measure the effect of a shock in the current period on the conditional variance in the subsequent period, and so facilitate comparison between models.

Definition 7.9 (News Impact Curve (NIC)). The news impact curve of an ARCH-family model is defined as the difference between the variance with a shock e_t and the variance with no shock ($e_t = 0$). The variance in all previous periods is set to the unconditional expectation of the variance, $\bar{\sigma}^2$,

$$n(e_t) = \sigma_{t+1}^2(e_t | \sigma_t^2 = \bar{\sigma}_t^2) \quad (7.36)$$

$$NIC(e_t) = n(e_t) - n(0). \quad (7.37)$$

Setting the variance in the current period to the unconditional variance has two consequences. First, it ensures that the NIC does not depend on the level of variance. Second, this choice for the lagged variance improves the comparison of linear and non-linear specifications (e.g., EGARCH).

News impact curves for ARCH and GARCH models only depend on the terms which include ε_t^2 .
GARCH(1,1)

$$n(e_t) = \omega + \alpha_1 \bar{\sigma}^2 e_t^2 + \beta_1 \bar{\sigma}^2 \quad (7.38)$$

$$NIC(e_t) = \alpha_1 \bar{\sigma}^2 e_t^2 \quad (7.39)$$

News impact curve are more complicated when models is not linear in ε_t^2 . For example, consider the NIC for a TARCH(1,1,1),

$$\sigma_t = \omega + \alpha_1 |\varepsilon_t| + \gamma_1 |\varepsilon_t| I_{[\varepsilon_t < 0]} + \beta_1 \sigma_{t-1}. \quad (7.40)$$

$$n(e_t) = \omega^2 + 2\omega(\alpha_1 + \gamma_1 I_{[\varepsilon_t < 0]})|\varepsilon_t| + 2\beta(\alpha_1 + \gamma_1 I_{[\varepsilon_t < 0]})|\varepsilon_t| \bar{\sigma} + \beta_1^2 \bar{\sigma}^2 + 2\omega\beta_1 \bar{\sigma} + (\alpha_1 + \gamma_1 I_{[\varepsilon_t < 0]})^2 \varepsilon_t^2 \quad (7.41)$$

$$NIC(e_t) = (\alpha_1 + \gamma_1 I_{[\varepsilon_t < 0]})^2 \varepsilon_t^2 + (2\omega + 2\beta_1 \bar{\sigma})(\alpha_1 + \gamma_1 I_{[\varepsilon_t < 0]})|\varepsilon_t| \quad (7.42)$$

While deriving explicit expressions for NICs can be tedious, practical implementation only requires computing the conditional variance for a shock of 0 ($n(0)$) and a set of shocks between -3 and 3 ($n(z)$ for $z \in (-3, 3)$). The difference between the conditional variance with a shock and the conditional variance without a shock is the NIC.

7.2.5.1 The S&P 500 and WTI

Figure 7.4 contains plots of the news impact curves for both the S&P 500 and WTI. When the models include asymmetries, the news impact curves are asymmetric and show a much larger response to negative shocks than to positive shocks, although the asymmetry is stronger in the volatility of the returns of the S&P 500 than it is in the volatility of WTI's returns.

7.3 Estimation and Inference

Consider a simple GARCH(1,1) specification,

$$r_t = \mu_t + \varepsilon_t \quad (7.43)$$

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2$$

$$\varepsilon_t = \sigma_t e_t$$

$$e_t \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$$

Since the errors are assumed to be conditionally i.i.d. normal⁹, maximum likelihood is a natural choice to estimate the unknown parameters, θ which contain both the mean and variance parameters. The normal likelihood for T independent variables is

$$f(\mathbf{r}; \theta) = \prod_{t=1}^T (2\pi\sigma_t^2)^{-\frac{1}{2}} \exp\left(-\frac{(r_t - \mu_t)^2}{2\sigma_t^2}\right) \quad (7.44)$$

and the normal log-likelihood function is

$$l(\theta; \mathbf{r}) = \sum_{t=1}^T -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma_t^2) - \frac{(r_t - \mu_t)^2}{2\sigma_t^2}. \quad (7.45)$$

If the mean is set to 0, the log-likelihood simplifies to

$$l(\theta; \mathbf{r}) = \sum_{t=1}^T -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma_t^2) - \frac{r_t^2}{2\sigma_t^2}, \quad (7.46)$$

and is maximized by solving the first order conditions.

$$\frac{\partial l(\theta; \mathbf{r})}{\partial \sigma_t^2} = \sum_{t=1}^T -\frac{1}{2\sigma_t^2} + \frac{r_t^2}{2\sigma_t^4} = 0, \quad (7.47)$$

which can be rewritten to provide some insight into the estimation of ARCH models,

$$\frac{\partial l(\theta; \mathbf{r})}{\partial \sigma_t^2} = \frac{1}{2} \sum_{t=1}^T \frac{1}{\sigma_t^2} \left(\frac{r_t^2}{\sigma_t^2} - 1 \right). \quad (7.48)$$

This expression clarifies that the parameters of the volatility are chosen to make $\left(\frac{r_t^2}{\sigma_t^2} - 1\right)$ as close to zero as possible.¹⁰ These first order conditions are not complete since ω , α_1 and β_1 , not σ_t^2 , are the parameters of a GARCH(1,1) model and

$$\frac{\partial l(\theta; \mathbf{r})}{\partial \theta_i} = \frac{\partial l(\theta; \mathbf{r})}{\partial \sigma_t^2} \frac{\partial \sigma_t^2}{\partial \theta_i} \quad (7.49)$$

⁹The use of conditional is to denote the dependence on σ_t^2 , which is in \mathcal{F}_{t-1} .

¹⁰If $E_{t-1} \left[\frac{r_t^2}{\sigma_t^2} - 1 \right] = 0$, and so the volatility is correctly specified, then the scores of the log-likelihood have expectation zero since

$$\begin{aligned} E \left[\frac{1}{\sigma_t^2} \left(\frac{r_t^2}{\sigma_t^2} - 1 \right) \right] &= E \left[E_{t-1} \left[\frac{1}{\sigma_t^2} \left(\frac{r_t^2}{\sigma_t^2} - 1 \right) \right] \right] \\ &= E \left[\frac{1}{\sigma_t^2} \left(E_{t-1} \left[\frac{r_t^2}{\sigma_t^2} - 1 \right] \right) \right] \\ &= E \left[\frac{1}{\sigma_t^2} (0) \right] \\ &= 0. \end{aligned}$$

The derivatives follow a recursive form not previously encountered,

$$\begin{aligned}\frac{\partial \sigma_t^2}{\partial \omega} &= 1 + \beta_1 \frac{\partial \sigma_{t-1}^2}{\partial \omega} \\ \frac{\partial \sigma_t^2}{\partial \alpha_1} &= \varepsilon_{t-1}^2 + \beta_1 \frac{\partial \sigma_{t-1}^2}{\partial \alpha_1} \\ \frac{\partial \sigma_t^2}{\partial \beta_1} &= \sigma_{t-1}^2 + \beta_1 \frac{\partial \sigma_{t-1}^2}{\partial \beta_1},\end{aligned}\tag{7.50}$$

although the recursion in the first order condition for ω can be removed noting that

$$\frac{\partial \sigma_t^2}{\partial \omega} = 1 + \beta_1 \frac{\partial \sigma_{t-1}^2}{\partial \omega} \approx \frac{1}{1 - \beta_1}.\tag{7.51}$$

Eqs. (7.49) – (7.51) provide the necessary formulas to implement the scores of the log-likelihood although they are not needed to estimate a GARCH model.¹¹

The use of the normal likelihood has one strong justification; estimates produced by maximizing the log-likelihood of a normal are *strongly consistent*. Strong consistency is a property of an estimator that ensures parameter estimates converge to the true parameters *even if the assumed conditional distribution is misspecified*. For example, in a standard GARCH(1,1), the parameter estimates would still converge to their true value if estimated with the normal likelihood as long as the volatility model was correctly specified. The intuition behind this result comes from the *generalized error*

$$\left(\frac{\varepsilon_t^2}{\sigma_t^2} - 1 \right).\tag{7.52}$$

Whenever $\sigma_t^2 = E_{t-1}[\varepsilon_t^2]$,

$$E \left[\left(\frac{\varepsilon_t^2}{\sigma_t^2} - 1 \right) \right] = E \left[\left(\frac{E_{t-1}[\varepsilon_t^2]}{\sigma_t^2} - 1 \right) \right] = E \left[\left(\frac{\sigma_t^2}{\sigma_t^2} - 1 \right) \right] = 0.\tag{7.53}$$

Thus, as long as the GARCH model nests the true DGP, the parameters are chosen to make the conditional expectation of the generalized error 0; these parameters correspond to those of the original DGP even if the conditional distribution is misspecified.¹² This is a unique property of the normal distribution and is not found in other common distributions.

7.3.1 Inference

Under some regularity conditions, parameters estimated by maximum likelihood are asymptotically normally distributed,

¹¹MATLAB and many other econometric packages are capable of estimating the derivatives using a numerical approximation that only requires the log-likelihood. Numerical derivatives use the definition of a derivative, $f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$ to approximate the derivative using $f'(x) \approx \frac{f(x+h) - f(x)}{h}$ for some small h .

¹²An assumption that a GARCH specification nests the DGP is extremely strong and likely wrong in most cases. However, the strong consistency property of the normal likelihood in volatility models justifies estimation of models where the standardized residuals are leptokurtotic and skewed.

$$\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \mathcal{I}^{-1}) \quad (7.54)$$

where

$$\mathcal{I} = -E \left[\frac{\partial^2 l(\theta_0; r_t)}{\partial \theta \partial \theta'} \right] \quad (7.55)$$

is the negative of the expected Hessian. The Hessian measures how much curvature there is in the log-likelihood at the optimum just like the second-derivative measures the rate-of-change in the rate-of-change of the function in a standard calculus problem. The sample analog estimator that averages the time-series of Hessian matrices computed at $\hat{\theta}$ is used to estimate \mathcal{I} ,

$$\hat{\mathcal{I}} = -T^{-1} \sum_{t=1}^T \frac{\partial^2 l(\hat{\theta}; r_t)}{\partial \theta \partial \theta'}. \quad (7.56)$$

Chapter 2 shows that the Information Matrix Equality (IME) generally holds for MLE problems, so that

$$\mathcal{I} = \mathcal{J} \quad (7.57)$$

where

$$\mathcal{J} = E \left[\frac{\partial l(r_t; \theta_0)}{\partial \theta} \frac{\partial l(r_t; \theta_0)}{\partial \theta'} \right] \quad (7.58)$$

is the covariance of the scores. The scores behave like errors in ML estimators and so large score variance indicate the parameters are difficult to estimate accurately. The estimator of \mathcal{J} is the sample analog averaging the outer-product of the scores evaluated at the estimated parameters,

$$\hat{\mathcal{J}} = T^{-1} \sum_{t=1}^T \frac{\partial l(\hat{\theta}; r_t)}{\partial \theta} \frac{\partial l(\hat{\theta}; r_t)}{\partial \theta'}. \quad (7.59)$$

The IME generally applies when the parameter estimates are *maximum likelihood estimates*, which requires that both the likelihood used in estimation is correct and that the specification for the conditional variance is general enough to nest the true process. When one specification is used for estimation (e.g., normal) but the data follow a different conditional distribution, these estimators are known as Quasi-Maximum Likelihood Estimators (QMLE), and the IME generally fails to hold. Under some regularity conditions, the estimated parameters are still asymptotically normal but with a different covariance,

$$\sqrt{T}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \mathcal{I}^{-1} \mathcal{J} \mathcal{I}^{-1}) \quad (7.60)$$

If the IME is valid, $\mathcal{I} = \mathcal{J}$ and so this covariance simplifies to the usual MLE variance estimator.

In most applications of ARCH models, the conditional distribution of shocks is decidedly not normal, and standardized residuals have excess kurtosis and are skewed. Bollerslev and Wooldridge (1992) were the first to show that the IME does not generally hold for GARCH models when the distribution is misspecified and the “sandwich” form

$$\hat{\mathcal{I}}^{-1} \hat{\mathcal{J}} \hat{\mathcal{I}}^{-1} \quad (7.61)$$

	WTI			
	ω	α_1	γ_1	β_1
Coefficient	0.031	0.030	0.055	0.942
Std. T-stat	3.62	4.03	7.67	102.94
Robust T-stat	1.85	2.31	4.45	49.66

	S&P 500			
	ω	α_1	γ_1	β_1
Coefficient	0.026	0.000	0.172	0.909
Std. T-stat	9.63	0.000	14.79	124.92
Robust T-stat	6.28	0.000	10.55	93.26

Table 7.3: Estimates from a TARCH(1,1,1) for the S&P 500 and WTI using alternative parameter covariance estimators.

of the covariance estimator is often referred to as the *Bollerslev-Wooldridge* covariance matrix or alternatively a robust covariance matrix. Standard Wald tests can be used to test hypotheses of interest, such as whether an asymmetric term is statistically significant, although likelihood ratio tests are not reliable since they do not have the usual χ_m^2 distribution.

7.3.1.1 The S&P 500 and WTI

A TARCH(1,1,1) models were estimated on both the S&P 500 and WTI returns to illustrate the differences between the MLE and the Bollerslev-Wooldridge (QMLE) covariance estimators. Table 7.3 contains the estimated parameters and t-stats using both the MLE covariance matrix and the Bollerslev-Wooldridge covariance matrix. The robust t-stats are substantially smaller than conventional ones, although conclusions about statistical significance are not affected except for ω in the WTI model. These changes are due to the heavy-tail in the standardized residuals, $\hat{e}_t = r_t - \hat{\mu}_t / \hat{\sigma}_t$, in these series.

7.3.1.2 Independence of the mean and variance parameters

Inference on the parameters of the ARCH model is still valid when using normal MLE or QMLE when the model for the mean is general enough to nest the true form. This property is important in practice since mean and variance parameters can be estimated separately without correcting the covariance matrix of the estimated parameters.¹³ This surprising feature of QMLE estimators employing a normal log-likelihood comes from the cross-partial derivative of the log-likelihood with respect to the mean and variance parameters,

¹³The estimated covariance for the mean should use a White covariance estimator. If the mean parameters are of particular interest, it may be more efficient to jointly estimate the parameters of the mean and volatility equations as a form of GLS (see Chapter 3).

$$l(\theta; r_t) = -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma_t^2) - \frac{(r_t - \mu_t)^2}{2\sigma_t^2}. \quad (7.62)$$

The first order condition is,

$$\frac{\partial l(\theta; \mathbf{r})}{\partial \mu_t} \frac{\partial \mu_t}{\partial \phi} = -\sum_{t=1}^T \frac{(r_t - \mu_t)}{\sigma_t^2} \frac{\partial \mu_t}{\partial \phi} \quad (7.63)$$

and the second order condition is

$$\frac{\partial^2 l(\theta; \mathbf{r})}{\partial \mu_t \partial \sigma_t^2} \frac{\partial \mu_t}{\partial \phi} \frac{\partial \sigma_t^2}{\partial \psi} = \sum_{t=1}^T \frac{(r_t - \mu_t)}{\sigma_t^4} \frac{\partial \mu_t}{\partial \phi} \frac{\partial \sigma_t^2}{\partial \psi} \quad (7.64)$$

where ϕ is a parameter of the conditional mean and ψ is a parameter of the conditional variance. For example, in a simple ARCH(1) model with a constant mean,

$$\begin{aligned} r_t &= \mu + \varepsilon_t \\ \sigma_t^2 &= \omega + \alpha_1 \varepsilon_{t-1}^2 \\ \varepsilon_t &= \sigma_t e_t \\ e_t &\stackrel{\text{i.i.d.}}{\sim} N(0, 1), \end{aligned} \quad (7.65)$$

$\phi = \mu$ and ψ can be either ω or α_1 . Taking expectations of the cross-partial,

$$\begin{aligned} \mathbb{E} \left[\frac{\partial^2 l(\theta; \mathbf{r})}{\partial \mu_t \partial \sigma_t^2} \frac{\partial \mu_t}{\partial \phi} \frac{\partial \sigma_t^2}{\partial \psi} \right] &= \mathbb{E} \left[\sum_{t=1}^T \frac{r_t - \mu_t}{\sigma_t^4} \frac{\partial \mu_t}{\partial \phi} \frac{\partial \sigma_t^2}{\partial \psi} \right] \\ &= \mathbb{E} \left[\mathbb{E}_{t-1} \left[\sum_{t=1}^T \frac{r_t - \mu_t}{\sigma_t^4} \frac{\partial \mu_t}{\partial \phi} \frac{\partial \sigma_t^2}{\partial \psi} \right] \right] \\ &= \mathbb{E} \left[\sum_{t=1}^T \frac{\mathbb{E}_{t-1} [r_t - \mu_t]}{\sigma_t^4} \frac{\partial \mu_t}{\partial \phi} \frac{\partial \sigma_t^2}{\partial \psi} \right] \\ &= \mathbb{E} \left[\sum_{t=1}^T \frac{0}{\sigma_t^4} \frac{\partial \mu_t}{\partial \phi} \frac{\partial \sigma_t^2}{\partial \psi} \right] \\ &= 0 \end{aligned} \quad (7.66)$$

it can be seen that the expectation of the cross derivative is 0. The intuition behind the result follows from noticing that when the mean model is correct for the conditional expectation of r_t , the term $r_t - \mu_t$ has conditional expectation 0 and knowledge of the variance is not needed. This argument is a similar one used to establish the validity of the OLS estimator when the errors are heteroskedastic.

7.4 GARCH-in-Mean

The GARCH-in-mean model (GiM) makes a significant change to the role of time-varying volatility by explicitly relating the level of volatility to the expected return (Engle, Lilien, and Robins, 1987). A simple GiM model can be specified as

$$\begin{aligned} r_t &= \mu + \delta \sigma_t^2 + \varepsilon_t \\ \sigma_t^2 &= \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \\ \varepsilon_t &= \sigma_t e_t \\ e_t &\stackrel{\text{i.i.d.}}{\sim} N(0, 1) \end{aligned} \tag{7.67}$$

although virtually any ARCH-family model could be used to model the conditional variance. The obvious difference between the GiM and a standard GARCH(1,1) is that the variance appears in the mean of the return. Note that the shock driving the changes in variance is not the mean return but still ε_{t-1}^2 , and so the ARCH portion of a GiM is unaffected. Other forms of the GiM model have been employed where the conditional standard deviation or the log of the conditional variance are used in the mean equation¹⁴,

$$r_t = \mu + \delta \sigma_t + \varepsilon_t \tag{7.68}$$

or

$$r_t = \mu + \delta \ln(\sigma_t^2) + \varepsilon_t \tag{7.69}$$

Because the variance appears in the mean equation for r_t , the mean and variance parameters cannot be separately estimated. Despite the apparent feedback, processes that follow a GiM are stationary as long as the variance process is stationary. The conditional variance (σ_t^2) in the conditional mean does not feedback into the conditional variance process and so behaves like an exogenous regressor.

7.4.1 The S&P 500

Standard asset pricing theory dictates that there is a risk-return trade-off. GARCH-in-mean models provide a natural method to test whether this is the case. Using the S&P 500 data, three GiM models were estimated (one for each transformation of the variance in the mean equation), and the results are presented in table 7.4. Based on these estimates, there does appear to be a trade-off between mean and variance and higher variances produce higher expected means, although the magnitude is economically small and the coefficients are only significant at the 10% level.

7.5 Alternative Distributional Assumptions

Despite the strengths of the assumption that the errors are conditionally normal – estimation is simple, and parameters are *strongly consistent* for the true parameters – GARCH models can be specified and estimated with alternative distributional assumptions. The motivation for using something other than the normal distribution is two-fold. First, a better approximation to the conditional distribution of

¹⁴The model for the conditional mean can be extended to include ARMA terms or any other predetermined regressor.

S&P 500 Garch-in-Mean Estimates							
	μ	δ	ω	α	γ	β	Log Lik.
σ^2	0.004 (0.753)	0.022 (0.074)	0.022 (0.000)	0.000 (0.999)	0.183 (0.000)	0.888 (0.000)	-6773.7
σ	-0.034 (0.304)	0.070 (0.087)	0.022 (0.000)	0.000 (0.999)	0.182 (0.000)	0.887 (0.000)	-6773.4
$\ln \sigma^2$	0.038 (0.027)	0.030 (0.126)	0.022 (0.000)	0.000 (0.999)	0.183 (0.000)	0.888 (0.000)	-6773.8

Table 7.4: GARCH-in-mean estimates for the S&P 500 series. δ measures the strength of the GARCH-in-mean, and so is the most interesting parameter. The volatility process was a standard GARCH(1,1). P-values are in parentheses.

the standardized returns may improve the precision of the volatility process parameter estimates and, in the case of MLE, the estimates will be fully efficient. Second, GARCH models are often used in applications where the choice of the assumed density is plays a larger role such as in Value-at-Risk estimation or option pricing.

Three distributions stand among the dozens that have been used to estimate the parameters of GARCH processes. The first is a standardized Student's t (to have a unit variance for any value v , see Bollerslev (1987)) with v degrees of freedom,

Standardized Student's t

$$f(r_t; \mu, \sigma_t^2, v) = \frac{\Gamma(\frac{v+1}{2})}{\Gamma(\frac{v}{2})} \frac{1}{\sqrt{\pi(v-2)}} \frac{1}{\sigma_t} \frac{1}{\left(1 + \frac{(r_t - \mu)^2}{\sigma_t^2(v-2)}\right)^{\frac{v+1}{2}}} \quad (7.70)$$

where $\Gamma(\cdot)$ is the gamma function.¹⁵ This distribution is always fat-tailed and produces a better fit than the normal for most asset return series. This distribution is only well defined if $v > 2$ since the variance of a Student's t with $v \leq 2$ is infinite. The second is the generalized error distribution (GED, see Nelson (1991)),

Generalized Error Distribution

$$f(r_t; \mu, \sigma_t^2, v) = \frac{v \exp\left(-\frac{1}{2} \left|\frac{r_t - \mu}{\sigma_t \lambda}\right|^v\right)}{\sigma_t \lambda 2^{\frac{v+1}{v}} \Gamma(\frac{1}{v})} \quad (7.71)$$

$$\lambda = \sqrt{\frac{2^{-\frac{2}{v}} \Gamma(\frac{1}{v})}{\Gamma(\frac{3}{v})}} \quad (7.72)$$

¹⁵The standardized Student's t differs from the usual Student's t so that it is necessary to scale data by $\sqrt{\frac{v}{v-2}}$ if using functions (such as the CDF) for the regular Student's t distribution.

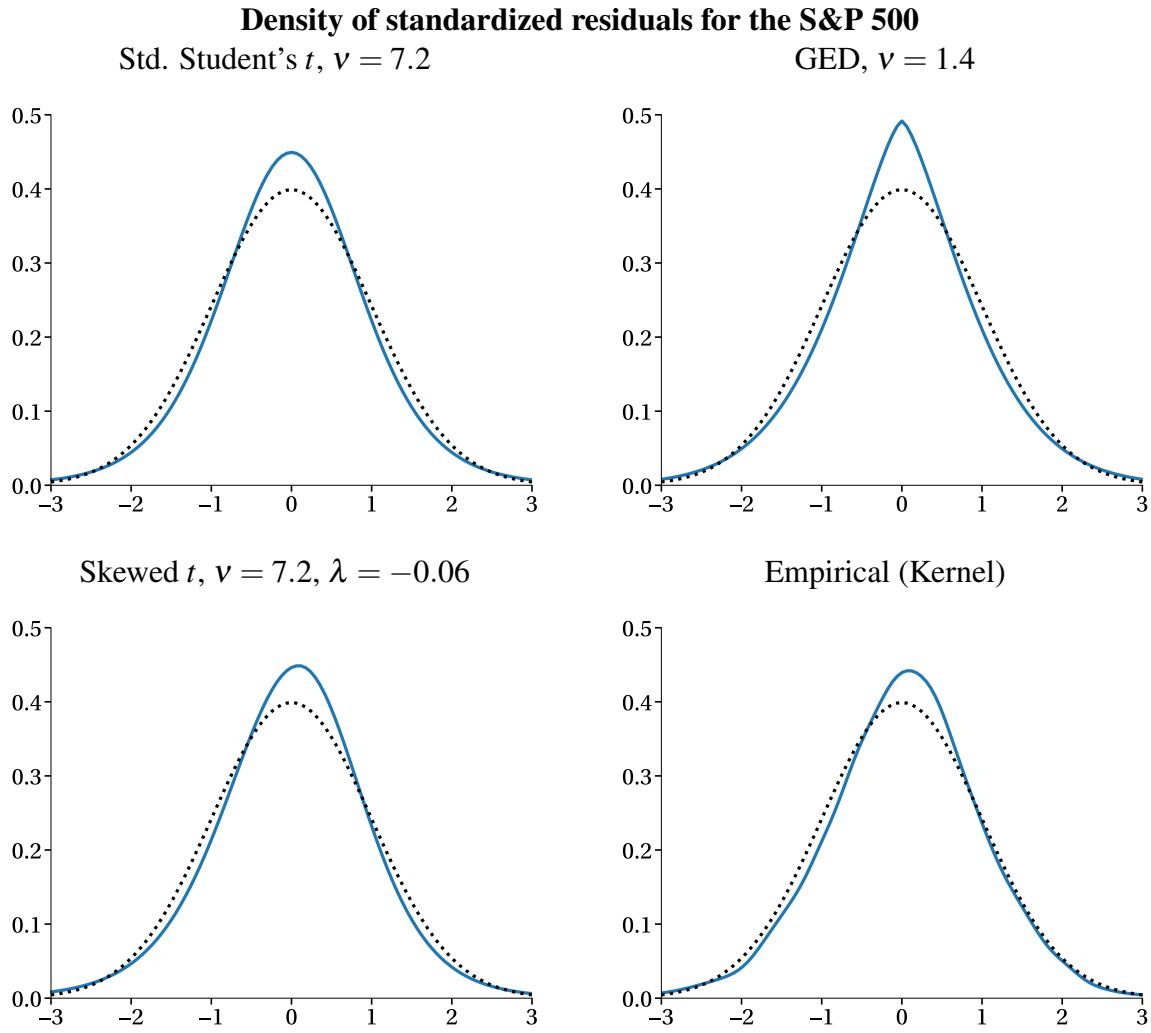


Figure 7.5: The four panels of this figure contain the estimated density for the S&P 500 and the density implied by the distributions: Student's t , GED, Hansen's Skew t and a kernel density plot of the standardized residuals, $\hat{\epsilon}_t = \epsilon_t / \hat{\sigma}_t$, along with the PDF of a normal (dotted line) for comparison. The shape parameters, v and λ , were jointly estimated with the variance parameters in the Student's t , GED, and skewed t .

which nests the normal when $v = 2$. The GED is fat-tailed when $v < 2$ and thin-tailed when $v > 2$. It is necessary that $v \geq 1$ to use the GED in volatility model estimation to ensure that variance is finite. The third useful distribution, introduced in Hansen (1994), extends the standardized Student's t to allow for skewness of returns

Hansen's skewed t

$$f(\varepsilon_t; \mu, \sigma_t, v, \lambda) = \begin{cases} bc \left(1 + \frac{1}{v-2} \left(\frac{b \left(\frac{r_t - \mu}{\sigma_t} \right) + a}{(1-\lambda)} \right)^2 \right)^{-(v+1)/2}, & \frac{r_t - \mu}{\sigma_t} < -a/b \\ bc \left(1 + \frac{1}{v-2} \left(\frac{b \left(\frac{r_t - \mu}{\sigma_t} \right) + a}{(1+\lambda)} \right)^2 \right)^{-(v+1)/2}, & \frac{r_t - \mu}{\sigma_t} \geq -a/b \end{cases} \quad (7.73)$$

where

$$a = 4\lambda c \left(\frac{v-2}{v-1} \right),$$

$$b = \sqrt{1 + 3\lambda^2 - a^2},$$

and

$$c = \frac{\Gamma \left(\frac{v+1}{2} \right)}{\sqrt{\pi(v-2)} \Gamma \left(\frac{v}{2} \right)}.$$

The two shape parameters, v and λ , control the kurtosis and the skewness, respectively.

These distributions may be better approximations to the actual distribution of the standardized residuals since they allow for kurtosis greater than that of the normal, an important empirical fact, and, in the case of the skewed t , skewness in the standardized returns. Chapter 8 applies these distributions in the context of Value-at-Risk and density forecasting.

7.5.1 Alternative Distribution in Practice

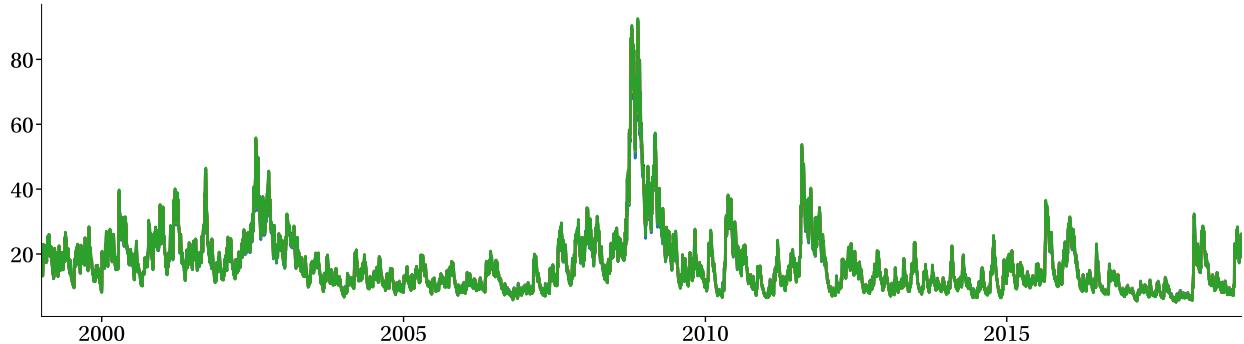
To explore the role of alternative distributional assumptions in the estimation of GARCH models, a TARCH(1,1,1) was fit to the S&P 500 returns using the conditional normal, the Student's t , the GED and Hansen's skewed t . Figure 7.5 contains the empirical density (constructed with a kernel) and the fit density of the three distributions. The shape parameters, v and λ , were jointly estimated with the conditional variance parameters. Figure 7.6 plots of the estimated conditional variance for both the S&P 500 and WTI using all four distributional assumptions. The most important aspect of this figure is that the fit variances are indistinguishable. This is a common finding: estimating models using alternative distributional assumptions produce little difference in the estimated parameters or the fitted conditional variances from the volatility model.¹⁶

7.6 Model Building

Since ARCH and GARCH models are similar to AR and ARMA models, the Box-Jenkins methodology is a natural way to approach the problem. The first step is to analyze the sample ACF and PACF

¹⁶While the volatilities are similar, the models do not fit the data equally well. The alternative distributions often provide a better fit as measured by the log-likelihood and provide a more accurate description of the probability in the tails of the distribution.

Conditional Variance and Distributional Assumptions
S&P 500 Annualized Volatility (TARCH(1,1,1))



WTI Annualized Volatility (TARCH(1,1,1))

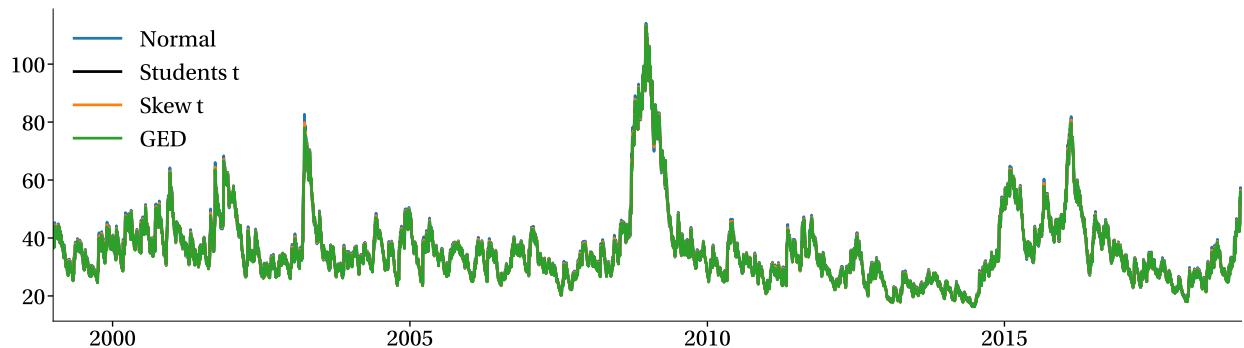


Figure 7.6: The choice of the distribution for the standardized innovation makes little difference to the fit variances or the estimated parameters in most models. The alternative distributions are more useful in application to Value-at-Risk and Density forecasting where the choice of density plays a more significant role.

of the *squared* returns, or if the model for the conditional mean is non-trivial, the sample ACF and PACF of the estimated residuals, $\hat{\epsilon}_t$, should be examined for heteroskedasticity. Figures 7.7 and 7.8 contains the ACF and PACF for the squared returns of the S&P 500 and WTI respectively. The models used in selecting the final model are reproduced in tables 7.5 and 7.6 respectively. Both selections began with a simple GARCH(1,1). The next step was to check if more lags were needed for either the squared innovation or the lagged variance by fitting a GARCH(2,1) and a GARCH(1,2) to each series. Neither of these meaningfully improved the fit, and a GARCH(1,1) was assumed to be sufficient to capture the symmetric dynamics.

The next step in model building is to examine whether the data exhibit any evidence of asymmetries using a GJR-GARCH(1,1,1). The asymmetry term was significant and so other forms of the GJR model were explored. All were found to provide little improvement in the fit. Once a GJR-GARCH(1,1,1) model was decided upon, a TARCH(1,1,1) was fit to examine whether evolution in variances or standard deviations was more appropriate for the data. Both series preferred the TARCH to the GJR-GARCH (compare the log-likelihoods), and the TARCH(1,1,1) was selected. In compar-

	α_1	α_2	γ_1	γ_2	β_1	β_2	Log Lik.
GARCH(1,1)	0.102 (0.000)				0.885 (0.000)		-6887.6
GARCH(1,2)	0.102 (0.000)				0.885 (0.000)	0.000 (0.999)	-6887.6
GARCH(2,1)	0.067 (0.003)	0.053 (0.066)			0.864 (0.000)		-6883.5
GJR-GARCH(1,1,1)	0.000 (0.999)		0.185 (0.000)		0.891 (0.000)		-6775.1
GJR-GARCH(1,2,1)	0.000 (0.999)		0.158 (0.000)	0.033 (0.460)	0.887 (0.000)		-6774.5
TARCH(1,1,1)*	0.000 (0.999)		0.172 (0.000)		0.909 (0.000)		-6751.9
TARCH(1,2,1)	0.000 (0.999)		0.165 (0.000)	0.009 (0.786)	0.908 (0.000)		-6751.8
TARCH(2,1,1)	0.000 (0.999)	0.003 (0.936)	0.171 (0.000)		0.907 (0.000)		-6751.9
EGARCH(1,0,1)	0.211 (0.000)				0.979 (0.000)		-6908.4
EGARCH(1,1,1)	0.136 (0.000)		-0.153 (0.000)		0.975 (0.000)		-6766.7
EGARCH(1,2,1)	0.129 (0.000)		-0.213 (0.000)	0.067 (0.045)	0.977 (0.000)		-6761.7
EGARCH(2,1,1)	0.020 (0.651)	0.131 (0.006)	-0.162 (0.000)		0.970 (0.000)		-6757.6

Table 7.5: The models estimated in selecting a final model for the conditional variance of the S&P 500 Index. * indicates the selected model.

ing alternative specifications, an EGARCH was fit and found to provide a good description of the data. In both cases, the EGARCH was expanded to include more lags of the shocks or lagged log volatility. The EGARCH did not improve over the TARCH for the S&P 500, and so the TARCH(1,1,1) was selected. The EGARCH did fit the WTI data better, and so the preferred model is an EGARCH(1,1,1), although a case could be made for the EGARCH(2,1,1) which provided a better fit. Overfitting is always a concern, and the opposite signs on α_1 and α_2 in the EGARCH(2,1,1) are suspicious.

7.6.0.1 Testing for (G)ARCH

Although conditional heteroskedasticity can often be identified by graphical inspection, a formal test of conditional homoskedasticity is also useful. The standard method to test for ARCH is to use the ARCH-LM test which is implemented as a regression of *squared* residuals on lagged squared residuals. The test directly exploits the AR representation of an ARCH process (Engle, 1982) and is computed as T times the R^2 ($LM = T \times R^2$) from the regression

$$\hat{\varepsilon}_t^2 = \phi_0 + \phi_1 \hat{\varepsilon}_{t-1}^2 + \dots + \phi_P \hat{\varepsilon}_{t-P}^2 + \eta_t. \quad (7.74)$$

The test statistic is asymptotically distributed χ_P^2 where $\hat{\varepsilon}_t$ are residuals constructed from the returns by subtracting the conditional mean. The null hypothesis is $H_0 : \phi_1 = \dots = \phi_P = 0$ which corresponds to no persistence in the conditional variance.

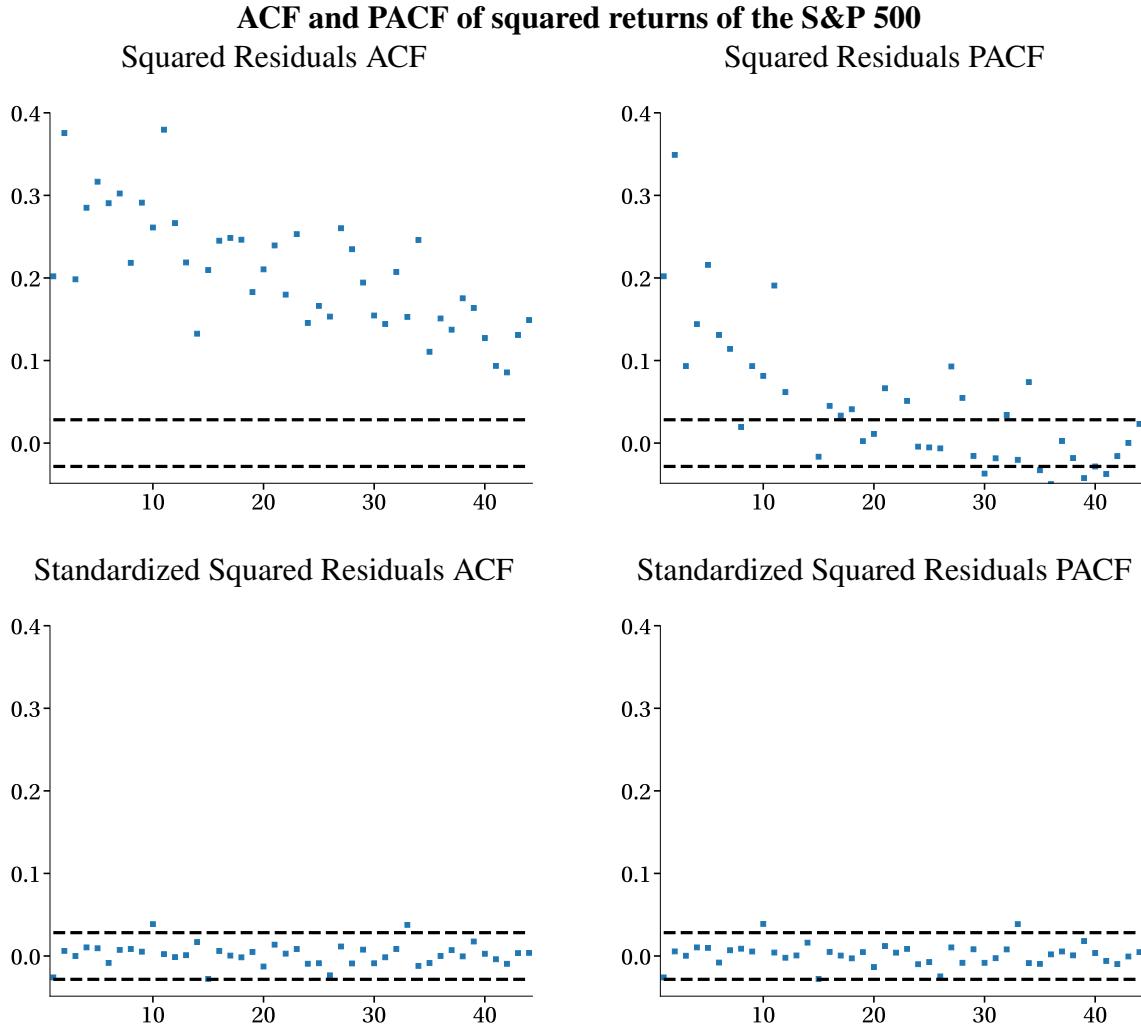


Figure 7.7: ACF and PACF of the squared returns for the S&P 500. The bottom two panels plot the ACF and PACF of the standardized squared residuals, $\hat{\varepsilon}_t^2 = \hat{\varepsilon}_t^2 / \hat{\sigma}_t^2$. The top panels indicate persistence through both the ACF and PACF. These plots suggest that a GARCH model is needed. The ACF and PACF of the standardized residuals are consistent with those of a white noise process.

7.7 Forecasting Volatility

Forecasting conditional variances with ARCH-family models ranges from simple for ARCH and GARCH processes to difficult for non-linear specifications. Consider the simple ARCH(1) process,

$$\begin{aligned} \varepsilon_t &= \sigma_t e_t \\ e_t &\stackrel{\text{i.i.d.}}{\sim} N(0, 1) \\ \sigma_t^2 &= \omega + \alpha_1 \varepsilon_{t-1}^2 \end{aligned} \tag{7.75}$$

Iterating forward, $\sigma_{t+1}^2 = \omega + \alpha_1 \varepsilon_t^2$, and taking conditional expectations, $E_t[\sigma_{t+1}^2] = E_t[\omega + \alpha_1 \varepsilon_t^2] =$

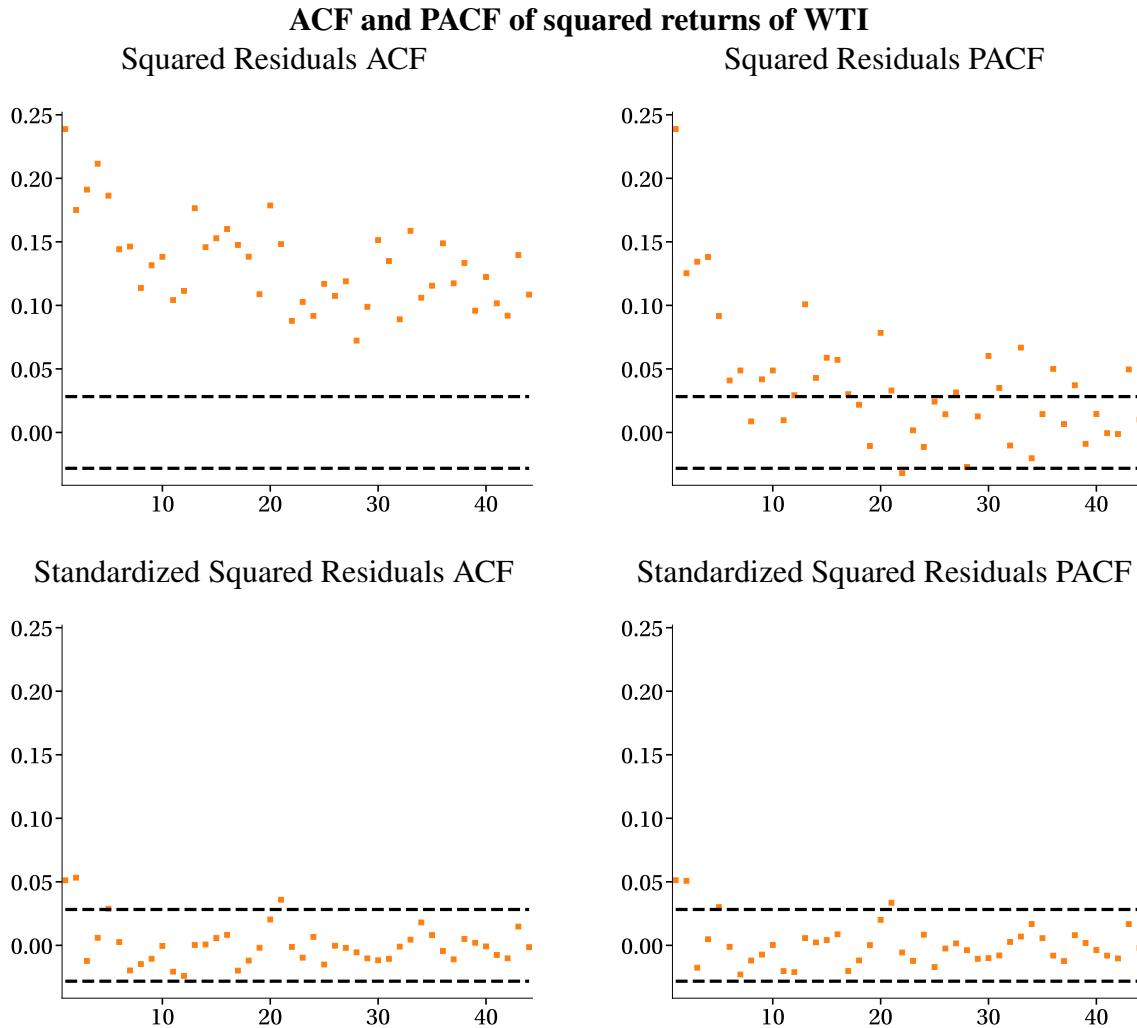


Figure 7.8: ACF and PACF of the squared returns for WTI. The bottom two panels plot the ACF and PACF of the standardized squared residuals, $\hat{\epsilon}_t^2 = \hat{\epsilon}_t^2 / \hat{\sigma}_t^2$. The top panels indicate persistence through both the ACF and PACF. These plots suggest that a GARCH model is needed. The ACF and PACF of the standardized residuals are consistent with those of a white noise process. When compared to the S&P 500 ACF and PACF, the ACF and PACF of the WTI returns indicate less persistence in volatility.

$\omega + \alpha_1 \varepsilon_t^2$ since all of these quantities are known at time t . This is a property common to *all* ARCH-family models: *the forecast of σ_{t+1}^2 is known at time t .*¹⁷

The 2-step ahead forecast follows from an application of the law of iterated expectations,

$$\begin{aligned} E_t[\sigma_{t+2}^2] &= E_t[\omega + \alpha_1 \varepsilon_{t+1}^2]. \\ &= \omega + \alpha_1 E_t[\varepsilon_{t+1}^2] \end{aligned} \tag{7.76}$$

¹⁷Not only is this property common to all ARCH-family members, but it is also the defining characteristic of an ARCH model.

	α_1	α_2	γ_1	γ_2	β_1	β_2	Log Lik.
GARCH(1,1)	0.059 (0.000)				0.934 (0.000)		-11030.1
GARCH(1,2)	0.075 (0.000)				0.585 (0.000)	0.331 (0.027)	-11027.4
GARCH(2,1)	0.059 (0.001)	0.000 (0.999)			0.934 (0.000)		-11030.1
GJR-GARCH(1,1,1)	0.026 (0.008)		0.049 (0.000)		0.945 (0.000)		-11011.9
GJR-GARCH(1,2,1)	0.026 (0.010)		0.049 (0.102)	0.000 (0.999)	0.945 (0.000)		-11011.9
TARCH(1,1,1)	0.030 (0.021)		0.055 (0.000)		0.942 (0.000)		-11005.6
TARCH(1,2,1)	0.030 (0.038)		0.055 (0.048)	0.000 (0.999)	0.942 (0.000)		-11005.6
TARCH(2,1,1)	0.030 (0.186)	0.000 (0.999)	0.055 (0.000)		0.942 (0.000)		-11005.6
EGARCH(1,0,1)	0.148 (0.000)				0.986 (0.000)		-11029.5
EGARCH(1,1,1) [†]	0.109 (0.000)		-0.050 (0.000)		0.990 (0.000)		-11000.6
EGARCH(1,2,1)	0.109 (0.000)		-0.056 (0.043)	0.006 (0.834)	0.990 (0.000)		-11000.5
EGARCH(2,1,1) [*]	0.195 (0.000)	-0.101 (0.019)	-0.049 (0.000)		0.992 (0.000)		-10994.4

Table 7.6: The models estimated in selecting a final model for the conditional variance of WTI. * indicates the selected model. † indicates a model that could be considered for model selection.

$$\begin{aligned} &= \omega + \alpha_1(\omega + \alpha_1 \varepsilon_t^2) \\ &= \omega + \alpha_1 \omega + \alpha_1^2 \varepsilon_t^2 \end{aligned}$$

The expression for an h -step ahead forecast can be constructed by repeated substitution and is given by

$$E_t[\sigma_{t+h}^2] = \sum_{i=0}^{h-1} \alpha_1^i \omega + \alpha_1^h \varepsilon_t^2. \quad (7.77)$$

An ARCH(1) is an AR(1), and this formula is identical to the expression for the multi-step forecast of an AR(1).

Forecasts from GARCH(1,1) models are constructed following the same steps. The one-step-ahead forecast is

$$\begin{aligned} E_t[\sigma_{t+1}^2] &= E_t[\omega + \alpha_1 \varepsilon_t^2 + \beta_1 \sigma_t^2] \\ &= \omega + \alpha_1 \varepsilon_t^2 + \beta_1 \sigma_t^2. \end{aligned} \quad (7.78)$$

The two-step-ahead forecast is

$$\begin{aligned}
E_t[\sigma_{t+2}^2] &= E_t[\omega + \alpha_1 \varepsilon_{t+1}^2 + \beta_1 \sigma_{t+1}^2] \\
&= \omega + \alpha_1 E_t[\varepsilon_{t+1}^2] + \beta_1 E_t[\sigma_{t+1}^2] \\
&= \omega + \alpha_1 E_t[e_{t+1}^2 \sigma_{t+1}^2] + \beta_1 E_t[\sigma_{t+1}^2] \\
&= \omega + \alpha_1 E_t[e_{t+1}^2] E_t[\sigma_{t+1}^2] + \beta_1 E_t[\sigma_{t+1}^2] \\
&= \omega + \alpha_1 \cdot 1 \cdot E_t[\sigma_{t+1}^2] + \beta_1 E_t[\sigma_{t+1}^2] \\
&= \omega + \alpha_1 E_t[\sigma_{t+1}^2] + \beta_1 E_t[\sigma_{t+1}^2] \\
&= \omega + (\alpha_1 + \beta_1) E_t[\sigma_{t+1}^2].
\end{aligned}$$

Substituting the one-step-ahead forecast, $E_t[\sigma_{t+1}^2]$, shows that the forecast only depends on time t information,

$$\begin{aligned}
E_t[\sigma_{t+2}^2] &= \omega + (\alpha_1 + \beta_1)(\omega + \alpha_1 \varepsilon_t^2 + \beta_1 \sigma_t^2) \\
&= \omega + (\alpha_1 + \beta_1)\omega + (\alpha_1 + \beta_1)\alpha_1 \varepsilon_t^2 + (\alpha_1 + \beta_1)\beta_1 \sigma_t^2.
\end{aligned} \tag{7.79}$$

Note that $E_t[\sigma_{t+3}^2] = \omega + (\alpha_1 + \beta_1)E_t[\sigma_{t+2}^2]$, and so

$$\begin{aligned}
E_t[\sigma_{t+3}^2] &= \omega + (\alpha_1 + \beta_1)(\omega + (\alpha_1 + \beta_1)\omega + (\alpha_1 + \beta_1)\alpha_1 \varepsilon_t^2 + (\alpha_1 + \beta_1)\beta_1 \sigma_t^2) \\
&= \omega + (\alpha_1 + \beta_1)\omega + (\alpha_1 + \beta_1)^2 \omega + (\alpha_1 + \beta_1)^2 \alpha_1 \varepsilon_t^2 + (\alpha_1 + \beta_1)^2 \beta_1 \sigma_t^2.
\end{aligned} \tag{7.80}$$

Repeated substitution reveals a pattern in the multi-step forecasts which is compactly expressed as

$$E_t[\sigma_{t+h}^2] = \sum_{i=0}^{h-1} (\alpha_1 + \beta_1)^i \omega + (\alpha_1 + \beta_1)^{h-1} (\alpha_1 \varepsilon_t^2 + \beta_1 \sigma_t^2). \tag{7.81}$$

Despite similarities to ARCH and GARCH models, forecasts from GJR-GARCH are complicated by the presence of the asymmetric term. If the expected value of the squared shock does not depend on the sign of the return, so that $E[e_t^2 | e_t < 0] = E[e_t^2 | e_t > 0] = 1$, then the probability that $e_{t-1} < 0$ appears in the forecasting formula. When the standardized residuals are normal (or any other symmetric distribution), then this probability is $\frac{1}{2}$. If the density is unknown, this probability must be estimated from the model residuals.

In the GJR-GARCH model, the one-step-ahead forecast is

$$E_t[\sigma_{t+1}^2] = \omega + \alpha_1 \varepsilon_t^2 + \alpha_1 \varepsilon_t^2 I_{[\varepsilon_t < 0]} + \beta_1 \sigma_t^2. \tag{7.82}$$

The two-step-ahead forecast is

$$E_t[\sigma_{t+2}^2] = \omega + \alpha_1 E_t[\varepsilon_{t+1}^2] + \alpha_1 E_t[\varepsilon_{t+1}^2 I_{[\varepsilon_{t+1} < 0]}] + \beta_1 E_t[\sigma_{t+1}^2] \tag{7.83}$$

$$= \omega + \alpha_1 E_t[\sigma_{t+1}^2] + \alpha_1 E_t[\varepsilon_{t+1}^2 | \varepsilon_{t+1} < 0] + \beta_1 E_t[\sigma_{t+1}^2]. \tag{7.84}$$

Assuming the residuals are conditionally normally distributed, then $E_t[\varepsilon_{t+1}^2 | \varepsilon_{t+1} < 0] = 0.5E[\sigma_{t+1}^2]$.

Multi-step forecasts from other models in the ARCH-family, particularly those that are not linear combinations of ε_t^2 , are nontrivial and generally do not have simple recursive formulas. For example, consider forecasting the variance from the simplest nonlinear ARCH-family member, a TARCH(1,0,0) model,

$$\sigma_t = \omega + \alpha_1 |\varepsilon_{t-1}| \quad (7.85)$$

As is *always* the case, the 1-step ahead forecast is known at time t ,

$$\begin{aligned} E_t[\sigma_{t+1}^2] &= E_t[(\omega + \alpha_1 |\varepsilon_t|)^2] \\ &= E_t[\omega^2 + 2\omega\alpha_1 |\varepsilon_t| + \alpha_1^2 \varepsilon_t^2] \\ &= \omega^2 + 2\omega\alpha_1 E_t[|\varepsilon_t|] + \alpha_1^2 E_t[\varepsilon_t^2] \\ &= \omega^2 + 2\omega\alpha_1 |\varepsilon_t| + \alpha_1^2 \varepsilon_t^2 \end{aligned} \quad (7.86)$$

The 2-step ahead forecast is more complicated and is given by

$$\begin{aligned} E_t[\sigma_{t+2}^2] &= E_t[(\omega + \alpha_1 |\varepsilon_{t+1}|)^2] \\ &= E_t[\omega^2 + 2\omega\alpha_1 |\varepsilon_{t+1}| + \alpha_1^2 \varepsilon_{t+1}^2] \\ &= \omega^2 + 2\omega\alpha_1 E_t[|\varepsilon_{t+1}|] + \alpha_1^2 E_t[\varepsilon_{t+1}^2] \\ &= \omega^2 + 2\omega\alpha_1 E_t[|\varepsilon_{t+1}| \sigma_{t+1}] + \alpha_1^2 E_t[\varepsilon_t^2 \sigma_{t+1}^2] \\ &= \omega^2 + 2\omega\alpha_1 E_t[|\varepsilon_{t+1}|] E_t[\sigma_{t+1}] + \alpha_1^2 E_t[\varepsilon_t^2] E_t[\sigma_{t+1}^2] \\ &= \omega^2 + 2\omega\alpha_1 E_t[|\varepsilon_{t+1}|](\omega + \alpha_1 |\varepsilon_t|) + \alpha_1^2 \cdot 1 \cdot (\omega^2 + 2\omega\alpha_1 |\varepsilon_t| + \alpha_1^2 \varepsilon_t^2) \end{aligned} \quad (7.87)$$

The challenge in multi-step ahead forecasting of a TARCH model arises since the forecast depends on more than $E_t[e_{t+h}^2] \equiv 1$. In the above example, the forecast depends on both $E_t[\varepsilon_{t+1}^2] = 1$ and $E_t[|\varepsilon_{t+1}|]$. When returns are normally distributed, $E_t[|\varepsilon_{t+1}|] = \sqrt{\frac{2}{\pi}}$, but if the driving innovations have a different distribution, this expectation will differ. The forecast is then, assuming the conditional distribution is normal,

$$E_t[\sigma_{t+2}^2] = \omega^2 + 2\omega\alpha_1 \sqrt{\frac{2}{\pi}}(\omega + \alpha_1 |\varepsilon_t|) + \alpha_1^2(\omega^2 + 2\omega\alpha_1 |\varepsilon_t| + \alpha_1^2 \varepsilon_t^2). \quad (7.88)$$

The difficulty in multi-step forecasting using “nonlinear” GARCH models – those which involve powers other than two – follows directly from Jensen’s inequality. In the case of TARCH,

$$E_t[\sigma_{t+h}^2]^2 \neq E_t[\sigma_{t+h}^2] \quad (7.89)$$

or in the general case of an arbitrary power,

$$E_t[\sigma_{t+h}^\delta]^{\frac{2}{\delta}} \neq E_t[\sigma_{t+h}^2]. \quad (7.90)$$

7.7.1 Evaluating Volatility Forecasts

The evaluation of volatility forecasts is similar to the evaluation of forecasts from conditional mean models with one caveat. In standard time series models, once time $t + h$ has arrived, the value of the variable being forecast is known. However, the value of σ_{t+h}^2 is always unknown in volatility model evaluation and so the realization must be replaced by a proxy. The standard choice is to use the squared return, r_t^2 . This proxy is reasonable if the squared conditional mean is small relative to the variance, a plausible assumption for high-frequency applications to daily or weekly returns. If using longer horizon measurements of returns, e.g., monthly returns, squared residuals ($\hat{\epsilon}_t^2$) estimated from a model for the conditional mean can be used instead. *Realized Variance*, $RV_t^{(m)}$, is an alternative choice to use as a proxy for the unobserved volatility (see section 7.8). Once a choice of proxy has been made, Generalized Mincer-Zarnowitz regressions can be used to assess forecast optimality,

$$r_{t+h}^2 - \hat{\sigma}_{t+h|t}^2 = \gamma_0 + \gamma_1 \hat{\sigma}_{t+h|t}^2 + \gamma_2 z_{1t} + \dots + \gamma_{K+1} z_{Kt} + \eta_t \quad (7.91)$$

where z_{jt} are any instruments known at time t . Common choices for z_{jt} include r_t^2 , $|r_t|$, r_t or indicator variables for the sign of the lagged return. The GMZ regression is testing one key property of a well-specified model: $E_t [r_{t+h}^2 - \hat{\sigma}_{t+h|t}^2] = 0$.

The GMZ regression in equation 7.91 has a heteroskedastic variance, and so a more accurate regression, GMZ-GLS, can be constructed as

$$\frac{r_{t+h}^2 - \hat{\sigma}_{t+h|t}^2}{\hat{\sigma}_{t+h|t}^2} = \gamma_0 \frac{1}{\hat{\sigma}_{t+h|t}^2} + \gamma_1 \frac{z_{1t}}{\hat{\sigma}_{t+h|t}^2} + \dots + \gamma_{K+1} \frac{z_{Kt}}{\hat{\sigma}_{t+h|t}^2} + v_t \quad (7.92)$$

$$\frac{r_{t+h}^2}{\hat{\sigma}_{t+h|t}^2} - 1 = \gamma_0 \frac{1}{\hat{\sigma}_{t+h|t}^2} + \gamma_1 \frac{z_{1t}}{\hat{\sigma}_{t+h|t}^2} + \dots + \gamma_{K+1} \frac{z_{Kt}}{\hat{\sigma}_{t+h|t}^2} + v_t \quad (7.93)$$

by dividing both sides by the time t forecast, $\hat{\sigma}_{t+h|t}^2$, where $v_t = \eta_t / \hat{\sigma}_{t+h|t}^2$. Equation 7.93 shows that the GMZ-GLS is a regression of the *generalized error* from a normal likelihood. If one were to use the Realized Variance as the proxy, the GMZ and GMZ-GLS regressions are

$$RV_{t+h} - \hat{\sigma}_{t+h|t}^2 = \gamma_0 + \gamma_1 \hat{\sigma}_{t+h|t}^2 + \gamma_2 z_{1t} + \dots + \gamma_{K+1} z_{Kt} + \eta_t \quad (7.94)$$

and

$$\frac{RV_{t+h} - \hat{\sigma}_{t+h|t}^2}{\hat{\sigma}_{t+h|t}^2} = \gamma_0 \frac{1}{\hat{\sigma}_{t+h|t}^2} + \gamma_1 \frac{\hat{\sigma}_{t+h|t}^2}{\hat{\sigma}_{t+h|t}^2} + \gamma_2 \frac{z_{1t}}{\hat{\sigma}_{t+h|t}^2} + \dots + \gamma_{K+1} \frac{z_{Kt}}{\hat{\sigma}_{t+h|t}^2} + \frac{\eta_t}{\hat{\sigma}_{t+h|t}^2}. \quad (7.95)$$

Diebold-Mariano tests can also be used to test the relative performance of two models. A loss function must be specified when implementing a DM test. Two natural choices for the loss function are MSE,

$$\left(r_{t+h}^2 - \hat{\sigma}_{t+h|t}^2 \right)^2 \quad (7.96)$$

and QML-loss (which is the *kernel* of the normal log-likelihood),

$$\left(\ln(\hat{\sigma}_{t+h|t}^2) + \frac{r_{t+h}^2}{\hat{\sigma}_{t+h|t}^2} \right). \quad (7.97)$$

The DM statistic is a t-test of the null $H_0 : E[\delta_t] = 0$ where

$$\delta_t = \left(r_{t+h}^2 - \hat{\sigma}_{A,t+h|t}^2 \right)^2 - \left(r_{t+h}^2 - \hat{\sigma}_{B,t+h|t}^2 \right)^2 \quad (7.98)$$

in the case of the MSE loss or

$$\delta_t = \left(\ln(\hat{\sigma}_{A,t+h|t}^2) + \frac{r_{t+h}^2}{\hat{\sigma}_{A,t+h|t}^2} \right) - \left(\ln(\hat{\sigma}_{B,t+h|t}^2) + \frac{r_{t+h}^2}{\hat{\sigma}_{B,t+h|t}^2} \right) \quad (7.99)$$

when using QML-loss. Statistically significant positive values of $\bar{\delta} = R^{-1} \sum_{r=1}^R \delta_r$ indicate that B is a better model than A while negative values indicate the opposite (recall R is used to denote the number of out-of-sample observations used to compute the DM statistic). The QML-loss is preferred since it is a “heteroskedasticity corrected” version of the MSE. For more on the evaluation of volatility forecasts using MZ regressions see Patton and Sheppard (2009).

7.8 Realized Variance

Realized Variance (RV) is a new econometric methodology for measuring the variance of asset returns. RV differs from ARCH-models since it does not require a specific model to measure the volatility. Realized Variance instead uses a nonparametric estimator of the variance that is computed *using ultra high-frequency data*.¹⁸

Suppose the log-price process, p_t , is continuously available and is driven by a standard Wiener process with a constant mean and variance,

$$dp_t = \mu dt + \sigma dW_t.$$

The coefficients are normalized so that the return during one day is the difference between p at two consecutive integers (e.g., $p_1 - p_0$ is the first day’s return). For the S&P 500 index, $\mu \approx .00031$ and $\sigma \approx .0125$, which correspond to 8% and 20% for the annualized mean and volatility, respectively.

Realized Variance is estimated by sampling p_t throughout the trading day. Suppose that prices on day t were sampled on a regular grid of $m+1$ points, $0, 1, \dots, m$ and let $p_{i,t}$ denote the i^{th} observation of the log price. The m -sample Realized Variance on day t is defined

$$RV_t^{(m)} = \sum_{i=1}^m (p_{i,t} - p_{i-1,t})^2 = \sum_{i=1}^m r_{i,t}^2. \quad (7.100)$$

Since the price process is a standard Brownian motion, each return is an i.i.d. normal random variable with mean μ/m and variance σ^2/m (or volatility of σ/\sqrt{m}). First, consider the expectation of $RV_t^{(m)}$,

¹⁸Realized Variance was invented somewhere between 1972 and 1997. However, its introduction to modern econometrics clearly dates to the late 1990s (Andersen and Bollerslev, 1998; Andersen, Bollerslev, Diebold, and Labys, 2003; Barndorff-Nielsen and Shephard, 2004).

$$\mathbb{E} \left[RV_t^{(m)} \right] = \mathbb{E} \left[\sum_{i=1}^m r_{i,t}^2 \right] = \mathbb{E} \left[\sum_{i=1}^m \left(\frac{\mu}{m} + \frac{\sigma}{\sqrt{m}} \varepsilon_{i,t} \right)^2 \right] \quad (7.101)$$

where $\varepsilon_{i,t}$ are i.i.d. standard normal random variables.

$$\begin{aligned} \mathbb{E} \left[RV_t^{(m)} \right] &= \mathbb{E} \left[\sum_{i=1}^m \left(\frac{\mu}{m} + \frac{\sigma}{\sqrt{m}} \varepsilon_{i,t} \right)^2 \right] \\ &= \mathbb{E} \left[\sum_{i=1}^m \frac{\mu^2}{m^2} + 2 \frac{\mu \sigma}{m^{\frac{3}{2}}} \varepsilon_{i,t} + \frac{\sigma^2}{m} \varepsilon_{i,t}^2 \right] \\ &= \mathbb{E} \left[\sum_{i=1}^m \frac{\mu^2}{m^2} \right] + \mathbb{E} \left[\sum_{i=1}^m 2 \frac{\mu \sigma}{m^{\frac{3}{2}}} \varepsilon_{i,t} \right] + \mathbb{E} \left[\sum_{i=1}^m \frac{\sigma^2}{m} \varepsilon_{i,t}^2 \right] \\ &= \frac{\mu^2}{m} + \sum_{i=1}^m 2 \frac{\mu \sigma}{m^{\frac{3}{2}}} \mathbb{E} [\varepsilon_{i,t}] + \sum_{i=1}^m \frac{\sigma^2}{m} \mathbb{E} [\varepsilon_{i,t}^2] \\ &= \frac{\mu^2}{m} + 2 \frac{\mu \sigma}{m^{\frac{3}{2}}} \sum_{i=1}^m \mathbb{E} [\varepsilon_{i,t}] + \frac{\sigma^2}{m} \sum_{i=1}^m \mathbb{E} [\varepsilon_{i,t}^2] \\ &= \frac{\mu^2}{m} + 2 \frac{\mu \sigma}{m^{\frac{3}{2}}} \sum_{i=1}^m 0 + \frac{\sigma^2}{m} \sum_{i=1}^m 1 \\ &= \frac{\mu^2}{m} + \frac{\sigma^2}{m} \\ &= \frac{\mu^2}{m} + \sigma^2 \end{aligned} \quad (7.102)$$

The expected value is nearly σ^2 , the variance, and it is asymptotically unbiased, $\lim_{m \rightarrow \infty} \mathbb{E} [RV_t^{(m)}] = \sigma^2$. The variance of $RV_t^{(m)}$ can be similarly computed,

$$\begin{aligned} \mathbb{V} \left[RV_t^{(m)} \right] &= \mathbb{V} \left[\sum_{i=1}^m \frac{\mu^2}{m^2} + 2 \frac{\mu \sigma}{m^{\frac{3}{2}}} \varepsilon_{i,t} + \frac{\sigma^2}{m} \varepsilon_{i,t}^2 \right] \\ &= \mathbb{V} \left[\sum_{i=1}^m \frac{\mu^2}{m^2} \right] + \mathbb{V} \left[\sum_{i=1}^m 2 \frac{\mu \sigma}{m^{\frac{3}{2}}} \varepsilon_{i,t} \right] + \mathbb{V} \left[\sum_{i=1}^m \frac{\sigma^2}{m} \varepsilon_{i,t}^2 \right] + 2 \text{Cov} \left[\sum_{i=1}^m \frac{\mu^2}{m^2}, \sum_{i=1}^m 2 \frac{\mu \sigma}{m^{\frac{3}{2}}} \varepsilon_{i,t} \right] \\ &\quad + 2 \text{Cov} \left[\sum_{i=1}^m \frac{\mu^2}{m^2}, \sum_{i=1}^m \frac{\sigma^2}{m} \varepsilon_{i,t}^2 \right] + 2 \text{Cov} \left[\sum_{i=1}^m 2 \frac{\mu \sigma}{m^{\frac{3}{2}}} \varepsilon_{i,t}, \sum_{i=1}^m \frac{\sigma^2}{m} \varepsilon_{i,t}^2 \right]. \end{aligned} \quad (7.103)$$

First, the variance and covariance terms that involve the mean term are all zero,

$$\mathbb{V} \left[\sum_{i=1}^m \frac{\mu^2}{m^2} \right] = \text{Cov} \left[\sum_{i=1}^m \frac{\mu^2}{m^2}, \sum_{i=1}^m 2 \frac{\mu \sigma}{m^{\frac{3}{2}}} \varepsilon_{i,t} \right] = \text{Cov} \left[\sum_{i=1}^m \frac{\mu^2}{m^2}, \sum_{i=1}^m \frac{\sigma^2}{m} \varepsilon_{i,t}^2 \right] = 0,$$

since $\frac{\mu^2}{m^2}$ is a constant. The remaining covariance term also has expectation 0 since $\varepsilon_{i,t}$ are i.i.d. standard normal and so have a skewness of 0,

$$\text{Cov} \left[\sum_{i=1}^m 2 \frac{\mu\sigma}{m^{\frac{3}{2}}} \varepsilon_{i,t}, \sum_{i=1}^m \frac{\sigma^2}{m} \varepsilon_{i,t}^2 \right] = 0$$

The other two terms can be shown to be (left as exercises)

$$\begin{aligned} \text{V} \left[\sum_{i=1}^m 2 \frac{\mu\sigma}{m^{\frac{3}{2}}} \varepsilon_{i,t} \right] &= 4 \frac{\mu^2 \sigma^2}{m^2} \\ \text{V} \left[\sum_{i=1}^m \frac{\sigma^2}{m} \varepsilon_{i,t}^2 \right] &= 2 \frac{\sigma^4}{m} \end{aligned}$$

and so

$$\text{V} \left[RV_t^{(m)} \right] = 4 \frac{\mu^2 \sigma^2}{m^2} + 2 \frac{\sigma^4}{m}. \quad (7.104)$$

The variance is decreasing as $m \rightarrow \infty$, $RV_t^{(m)}$ is asymptotically unbiased, and so $RV_t^{(m)}$ is a consistent estimator of σ^2 .

In the empirically realistic case where the price process has a time-varying drift and stochastic volatility,

$$dp_t = \mu_t dt + \sigma_t dW_t,$$

$RV_t^{(m)}$ is a consistent estimator of the *integrated variance*,

$$\lim_{m \rightarrow \infty} RV_t^{(m)} \xrightarrow{p} \int_t^{t+1} \sigma_s^2 ds. \quad (7.105)$$

The integrated variance measures the average variance of the measurement interval, usually a day.

If the price process contains jumps, $RV_t^{(m)}$ is still a consistent estimator although its limit is the *quadratic variation* rather than the integrated variance, and so

$$\lim_{m \rightarrow \infty} RV_t^{(m)} \xrightarrow{p} \int_t^{t+1} \sigma_s^2 ds + \sum_{s \leq t} \Delta J_s^2. \quad (7.106)$$

where $\sum_{s \leq t} \Delta J_s^2$ is the sum of the squared jumps if any. Similar results hold if the price process exhibits leverage (instantaneous correlation between the price and the variance). The two conditions for $RV_t^{(m)}$ to be a reasonable method to estimate the integrated variance on day t are essentially that the price process, p_t , is arbitrage-free and that the efficient price is observable. Empirical evidence suggests that prices of liquid asset are compatible with the first condition. The second condition is violated since assets trade at either the best bid or best ask price – neither of which is the efficient price.

7.8.1 Implementing Realized Variance

In practice, naïve implementations of Realized Variance do not perform well. The most pronounced challenge is that observed prices are contaminated by noise; there is no single price, and traded prices are only observed at the bid and the ask. This feature of asset price transactions produces bid-ask bounce where consecutive prices oscillate between the two. Consider a simple model of bid-ask bounce where returns are computed as the log difference in observed prices composed of the true (unobserved) efficient prices, $p_{i,t}^*$, contaminated by an independent mean zero shock, $v_{i,t}$,

$$p_{i,t} = p_{i,t}^* + v_{i,t}.$$

The shock $v_{i,t}$ captures the difference between the efficient price and the observed prices which are always on the bid or ask price.

The i^{th} observed return, $r_{i,t}$ can be decomposed into the actual (unobserved) return $r_{i,t}^*$ and an independent noise term $\eta_{i,t} = v_{i,t} - v_{i-1,t}$,

$$\begin{aligned} p_{i,t} - p_{i-1,t} &= (p_{i,t}^* + v_{i,t}) - (p_{i-1,t}^* + v_{i-1,t}) \\ p_{i,t} - p_{i-1,t} &= (p_{i,t}^* - p_{i-1,t}^*) + (v_{i,t} - v_{i-1,t}) \\ r_{i,t} &= r_{i,t}^* + \eta_{i,t} \end{aligned} \tag{7.107}$$

The error in the observed return process, $\eta_{i,t} = v_{i,t} - v_{i-1,t}$, is a MA(1) and so is serially correlated.

Computing the RV from returns contaminated by noise has an unambiguous effect on Realized Variance; RV is biased upward.

$$\begin{aligned} RV_t^{(m)} &= \sum_{i=1}^m r_{i,t}^2 \\ &= \sum_{i=1}^m (r_{i,t}^* + \eta_{i,t})^2 \\ &= \sum_{i=1}^m r_{i,t}^{*2} + 2r_{i,t}^* \eta_{i,t} + \eta_{i,t}^2 \\ &\approx \widehat{RV}_t + m\tau^2 \end{aligned} \tag{7.108}$$

where τ^2 is the variance of $\eta_{i,t}$ and \widehat{RV}_t is the Realized Variance that would be computed if the efficient returns could be observed. The bias is increasing in the number of samples (m) and can be substantial for assets with large bid-ask spreads.

The simplest “solution” to the bias is to avoid the issue using *sparse sampling*, i.e., not using all of the observed prices. The noise imposes limits on m to ensure that the bias is small relative to the integrated variance. In practice the maximum m is always much higher than 1 – a single open-to-close return – and is typically somewhere between 13 (30-minute returns on a stock listed on the NYSE) and 390 (1-minute returns), and so even when $RV_t^{(m)}$ is not consistent, it is still a better proxy, often substantially, for the latent variance on day t than r_t^2 (the “1-sample Realized Variance”, see Bandi and Russell (2008)). The signal-to-noise ratio (which measures the ratio of useful information to pure

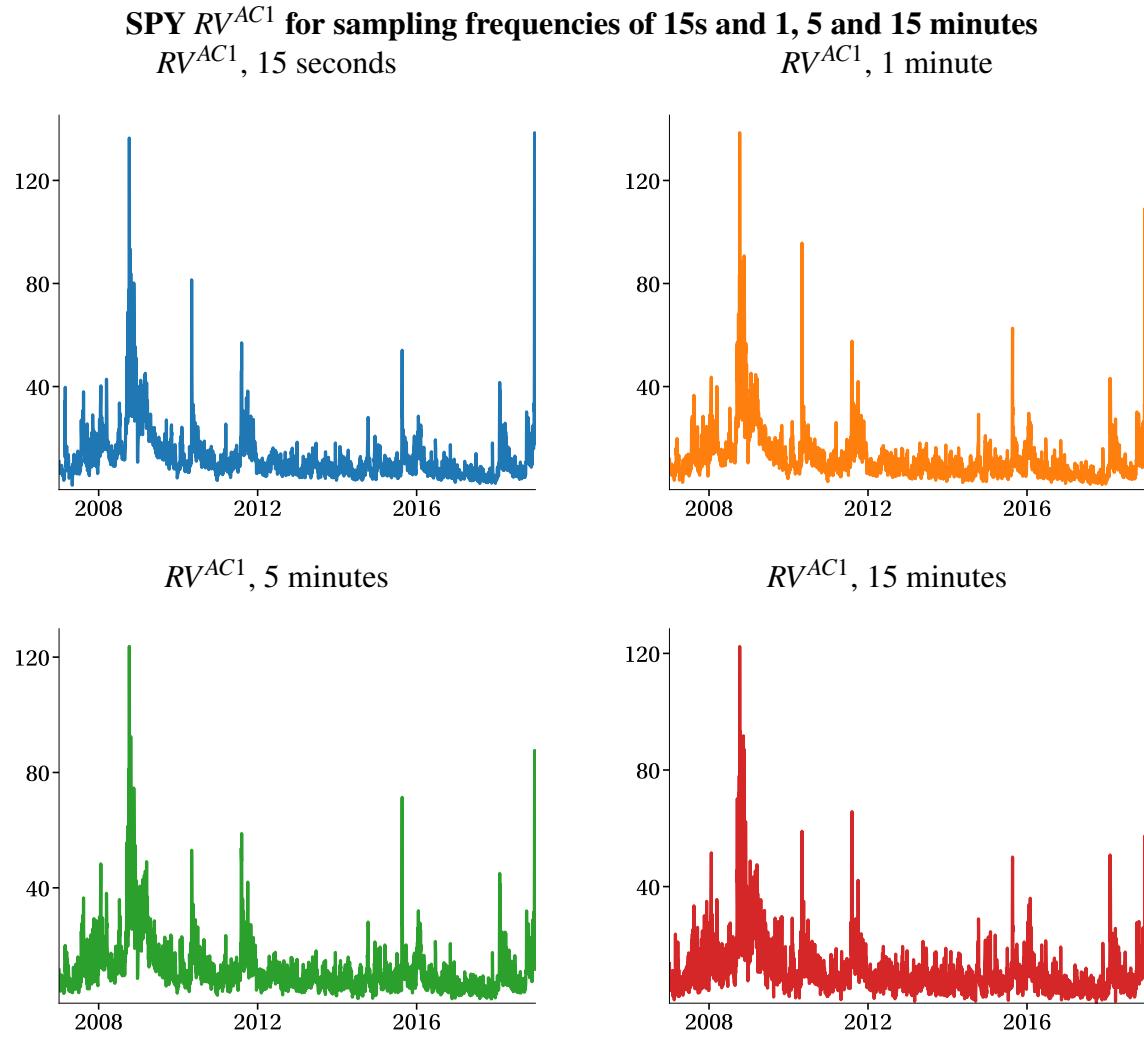


Figure 7.9: The four panels of this figure contain a noise-robust version Realized Variance, RV^{AC1} , for every day the market was open from January 2007 until December 2018 transformed into annualized volatility. The 15-second RV^{AC1} is better behaved than the 15-second RV .

noise) is approximately 1 for RV but is between .05 and .1 for r_t^2 . In other words, RV is 10-20 times more precise than squared daily returns (Andersen and Bollerslev, 1998).

Another simple and effective method is to filter the data using an MA(1). Transaction data contain a strong negative MA due to bid-ask bounce, and so RV computed using the errors ($\hat{\epsilon}_{i,t}$) from a model,

$$r_{i,t} = \theta \hat{\epsilon}_{i-1,t} + \epsilon_{i,t} \quad (7.109)$$

eliminates much of the bias. A better method to remove the bias is to use an estimator known as RV^{AC1} which is similar to a Newey-West estimator.

$$RV_t^{AC1(m)} = \sum_{i=1}^m r_{i,t}^2 + 2 \sum_{i=2}^m r_{i,t} r_{i-1,t} \quad (7.110)$$

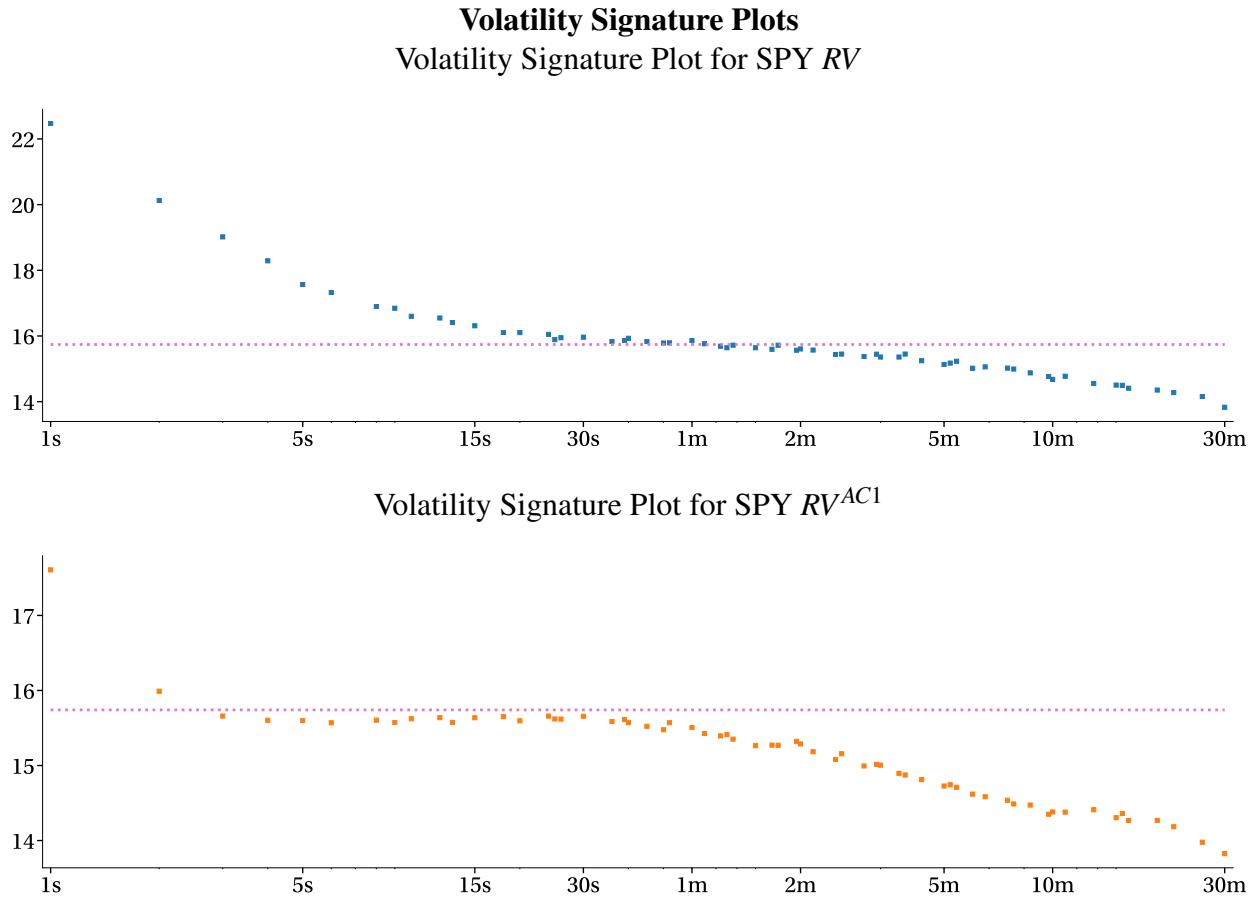


Figure 7.10: The volatility signature plot for the RV shows a clear trend. Based on visual inspection, it would be difficult to justify sampling more frequently than 30 seconds. Unlike the volatility signature plot of the RV , the signature plot of RV^{AC1} does not monotonically increase with the sampling frequency except when sampling every second, and the range of the values is considerably smaller than in the RV signature plot.

In the case of a constant drift, constant volatility Brownian motion subject to bid-ask bounce, this estimator can be shown to be unbiased, although it is not consistent in large samples. A more general class of estimators that use a kernel structure that can be tuned to match the characteristics of specific asset prices and which are consistent as $m \rightarrow \infty$ even in the presence of noise has been introduced in Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008).¹⁹

Another problem for Realized Variance is that prices are not available at regular intervals. Fortunately, this issue has a simple solution: *last price interpolation*. Last price interpolation sets the price at time t to the last observed price p_τ where τ is the largest time index less where $\tau \leq t$. Other interpolation schemes produce bias in RV . Consider, for example, linear interpolation which sets prices at time- t price to $p_t = wp_{\tau_1} + (1 - w)p_{\tau_2}$ where τ_1 is the time subscript of the last observed

¹⁹The Newey-West estimator is a particular implementation of a broad class of estimators known as kernel variance estimators. They all share the property that they are weighted sums of autocovariances where a kernel function determines the weights.

price before t and τ_2 is the time subscript of the first price after time t , and the interpolation weight is $w = (\tau_2 - t)/(\tau_2 - \tau_1)$. The averaging of prices in linear interpolation – which effectively produces a smoother price path than the efficient price path – produces a notable downward bias in RV .

Finally, most markets do not operate 24 hours a day, and RV cannot be computed when markets are closed. The standard procedure is to augment high-frequency returns with the squared close-to-open return to construct an estimate of the total variance. The close-to-close (CtC) RV is then defined

$$RV_{\text{CtC},t}^{(m)} = r_{\text{CtO},t}^2 + RV_t^{(m)} \quad (7.111)$$

where $r_{\text{CtO},t}^2$ is the return between the close on day $t - 1$ and the market open on day t . Since the overnight return is not measured frequently, the adjusted RV must be treated as a random variable (and not an observable). An improved method to handle the overnight return has been proposed in Hansen and Lunde (2005) and Hansen and Lunde (2006) which weighs the overnight squared return by λ_1 and the daily Realized Variance by λ_2 to produce an estimator with a lower mean-square error,

$$\widetilde{RV}_{\text{CtC},t}^{(m)} = \lambda_1 r_{\text{CtO},t}^2 + \lambda_2 RV_t^{(m)}.$$

7.8.2 Modeling RV

If RV is observable, then it can be modeled using standard time series tools such as ARMA models. This approach has been widely used in the academic literature although there are issues in treating the RV “as-if” it is the variance. If RV has measurement error, then parameter estimates in ARMA models suffer from an errors-in-variables problem, and the estimated coefficient are biased (see chapter 4). Corsi (2009) proposed the *heterogeneous autoregression* (HAR) as a simple method to capture the dynamics in RV in a parsimonious model. The standard HAR models the RV as a function of the RV in the previous day, the average RV over the previous week, and the average RV over the previous month (22 days). The HAR in levels is then

$$RV_t = \phi_0 + \phi_1 RV_{t-1} + \phi_5 \overline{RV}_{t-5} + \phi_2 2\overline{RV}_{t-22} + \varepsilon_t \quad (7.112)$$

where $\overline{RV}_{t-5} = \frac{1}{5} \sum_{i=1}^5 RV_{t-i}$ and $\overline{RV}_{t-22} = \frac{1}{22} \sum_{i=1}^2 2RV_{t-i}$ (suppressing the (m) terms). The HAR is also commonly estimated in logs,

$$\ln RV_t = \phi_0 + \phi_1 \ln RV_{t-1} + \phi_5 \ln \overline{RV}_{t-5} + \phi_2 2 \ln \overline{RV}_{t-22} + \varepsilon_t. \quad (7.113)$$

HARs are technically AR(22) models with many parameter restrictions. These restrictions maintain parsimony while allowing HARs to capture both the high degree of persistence in volatility (through the 22-day moving average) and short term dynamics (through the 1-day and 5-day terms).

The alternative is to model RV using ARCH-family models, which can be interpreted as multiplicative error models for *any* non-negative process, not only squared returns (Engle, 2002a).²⁰ Standard statistical software can be used to model RV as an ARCH process by defining $\tilde{r}_t = \text{sign}(r_t)\sqrt{RV_t}$ where $\text{sign}(r_t)$ is 1 if the end-of-day return is positive or -1 otherwise. The transformed RV , \tilde{r}_t , is the signed square root of the Realized Variance on day t . Any ARCH-family model can be applied to these

²⁰ARCH-family models have, for example, been successfully applied to both durations (time between trades) and hazards (number of trades in an interval of time), two non-negative processes.

transformed values. For example, when modeling the variance evolution as a GJR-GARCH(1,1,1) process,

$$\sigma_t^2 = \omega + \alpha_1 \tilde{r}_{t-1}^2 + \gamma_1 \tilde{r}_{t-1}^2 I_{[\tilde{r}_{t-1} < 0]} + \beta_1 \sigma_{t-1}^2 \quad (7.114)$$

which is equivalently expressed in terms of Realized Variance as

$$\sigma_t^2 = \omega + \alpha_1 RV_{t-1} + \gamma_1 RV_{t-1} I_{[r_{t-1} < 0]} + \beta_1 \sigma_{t-1}^2. \quad (7.115)$$

Maximum likelihood estimation, assuming normally distributed errors, can be used to estimate the parameters of this model. This procedure solves the errors-in-variables problem present when RV is treated as observable and facilitates modeling RV using standard software. Inference and the method to build a model are unaffected by the change from end-of-day returns to the transformed RV .

7.8.3 Realized Variance of the S&P 500

Returns on S&P 500 Depository Receipts, known as SPiDeRs (NYSEARCA:SPY) is used to illustrate the gains and pitfalls of RV . Price data was taken from TAQ and includes every transaction between January 2007 until December 2018, a total of 3,020 days. SPDRs track the S&P 500 and are among the most liquid assets in the U.S. market with an average volume of 150 million shares per day. There were more than 100,000 trades on a typical day throughout the sample, which is more than 4 per second. TAQ data contain errors, and observations were filtered by removing the prices outside the daily high or low from an audited database. Only trade prices that occurred during the usual trading hours of 9:30 – 16:00 were retained.

The primary tool for examining different Realized Variance estimators is the volatility signature plot.

Definition 7.10 (Volatility Signature Plot). The volatility signature plot displays the time-series average of Realized Variance

$$\overline{RV}_t^{(m)} = T^{-1} \sum_{t=1}^T RV_t^{(m)}$$

as a function of the number of samples, m . An equivalent representation displays the amount of time, whether in calendar time or tick time (number of trades between observations) along the X-axis.

Figures 7.11 and 7.9 contain plots of the *annualized volatility* constructed from the RV and RV^{AC1} . The estimates have been annualized to facilitate interpretation. Figures 7.11 shows that the 15-second RV is larger than the RV sampled at 1, 5 or 15 minutes and that the 1 and 5 minute RV are less noisy than the 15-minute RV . These plots provide some evidence that sampling more frequently than 15 minutes may be desirable. The two figures show that there is a reduction in the scale of the 15-second RV^{AC1} relative to the 15-second RV . The 15-second RV is heavily influenced by the noise in the data (bid-ask bounce) while the RV^{AC1} is less affected.

Figures 7.10 and 7.10 contain the annualized volatility signature plot for RV and RV^{AC1} , respectively. The dashed horizontal line depicts the volatility computed using the standard variance estimator computed from open-to-close returns. There is a striking difference between the two figures. The RV volatility signature plot diverges when sampling more frequently than 30 seconds while the RV^{AC1} plot is flat except at the highest sample frequency. RV^{AC1} appears to allow sampling every 5

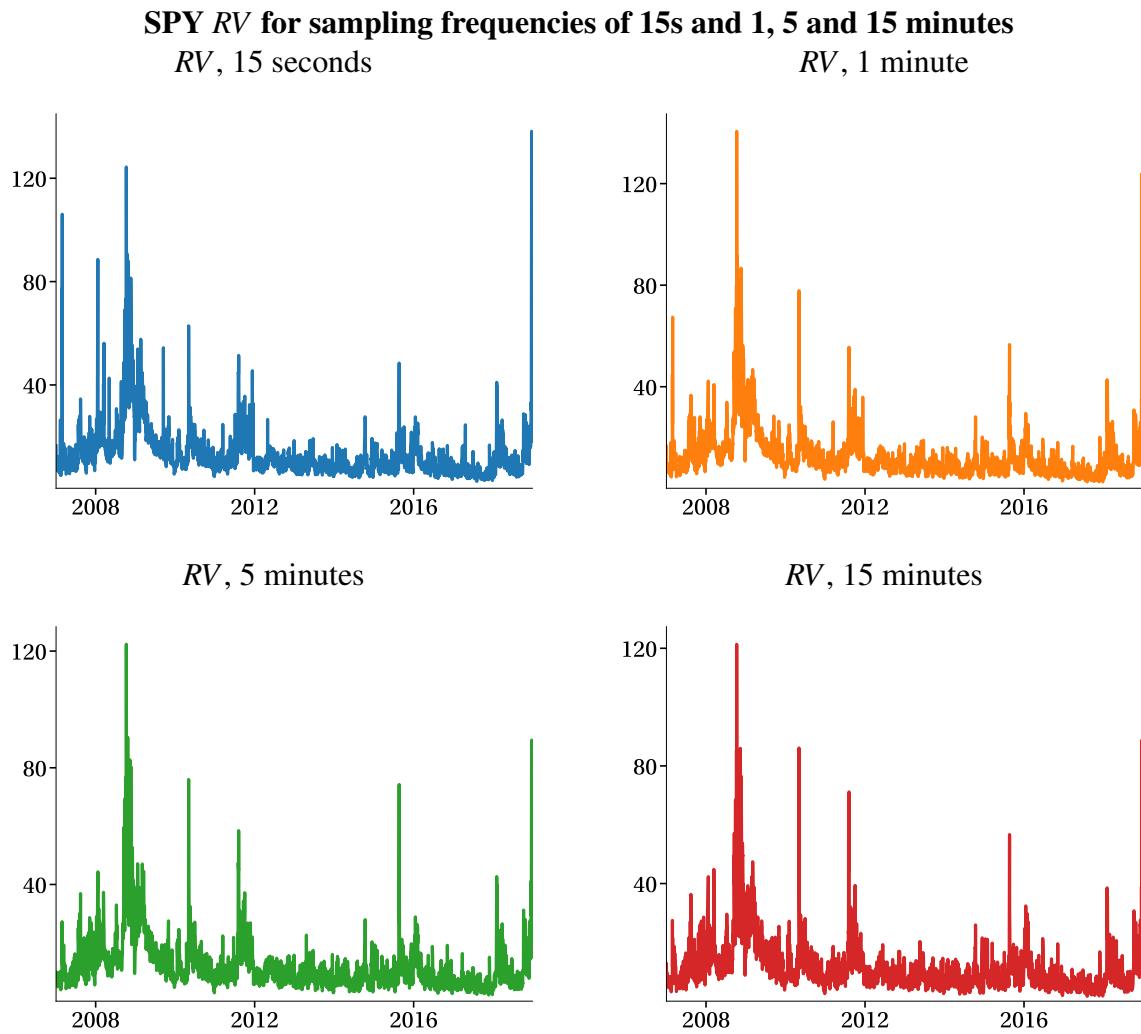


Figure 7.11: The four panels of this figure contain the Realized Variance for every day the market was open from January 2007 until December 2018. The estimated RV have been transformed into annualized volatility ($\sqrt{252 \cdot RV_t^{(m)}}$). While these plots appear superficially similar, the 1- and 5-minute RV are the most precise and the 15-second RV is biased upward.

seconds – 6 times more frequently than RV . This is a common finding when comparing RV^{AC1} to RV across a wide range of asset price data.

7.9 Implied Volatility and VIX

Implied volatility differs from other measures in that it is both market-based and forward-looking. Implied volatility was originally conceived as the “solution” to the Black-Scholes options pricing formula where all values except the volatility are observable. Recall that the Black-Scholes formula is derived from an assumption that stock prices follow a geometric Brownian motion plus drift,

$$dS_t = \mu S_t dt + \sigma S_t dW_t \quad (7.116)$$

where S_t is the time t stock price, μ is the drift, σ is the (constant) volatility, and dW_t is a Wiener process. Under some additional assumptions sufficient to ensure no arbitrage, the price of a call option can be shown to be

$$\begin{aligned} C_t(T, K) &= S_t \Phi(d_1) + K e^{-rT} \Phi(d_2) \\ d_1 &= \frac{\ln(S_t/K) + (r + \sigma^2/2) T}{\sigma \sqrt{T}} \\ d_2 &= \frac{\ln(S_t/K) + (r - \sigma^2/2) T}{\sigma \sqrt{T}} \end{aligned} \quad (7.117)$$

where K is the strike price, T is the time to maturity, reported in years, r is the risk-free interest rate, and $\Phi(\cdot)$ is the normal CDF. The price of a call option is monotonic in the volatility, and so the formula can be inverted to express the volatility as a function of the call price and other observables. The implied volatility,

$$\sigma_t^{\text{Implied}} = g(C_t(T, K), S_t, K, T, r), \quad (7.118)$$

is the expected volatility between t and T under the risk-neutral measure (which is the same as under the physical when volatility is constant).²¹

7.9.1 The smile

When computing the Black-Scholes implied volatility across a range of strikes, the volatility usually resembles a “smile” (higher IV for out-of-the-money options than in the money) or “smirk” (higher IV for out-of-the-money puts). This pattern emerges since asset returns are heavy-tailed (“smile”) and skewed (“smirk”). The BSIV is derived under an assumption that the asset price follows a *geometric Brownian motion* so that the log returns are assumed to be normal. The smile reflects misspecification of the model underlying the Black-Scholes option pricing formula. Figure 7.12 shows the smile in the BSIV for SPY out-of-the-month options on January 15, 2017. The x-axis rescaled from the strike price to *moneyness* by dividing the strike by the spot price. The current spot price is 100, smaller values indicate strikes below the current price (out-of-the-money puts), and positive values are strikes above the current price (out-of-the-money calls).

7.9.2 Model-Free Volatility

B-S implied volatility suffers from three key issues:

- Derived under constant volatility: The returns on most asset prices exhibit conditional heteroskedasticity, and time-variation in the volatility of returns generates heavy tails which increases the probability of a large asset price change.

²¹The implied volatility is computed by numerically inverting the B-S pricing formula, or using some other approximation..

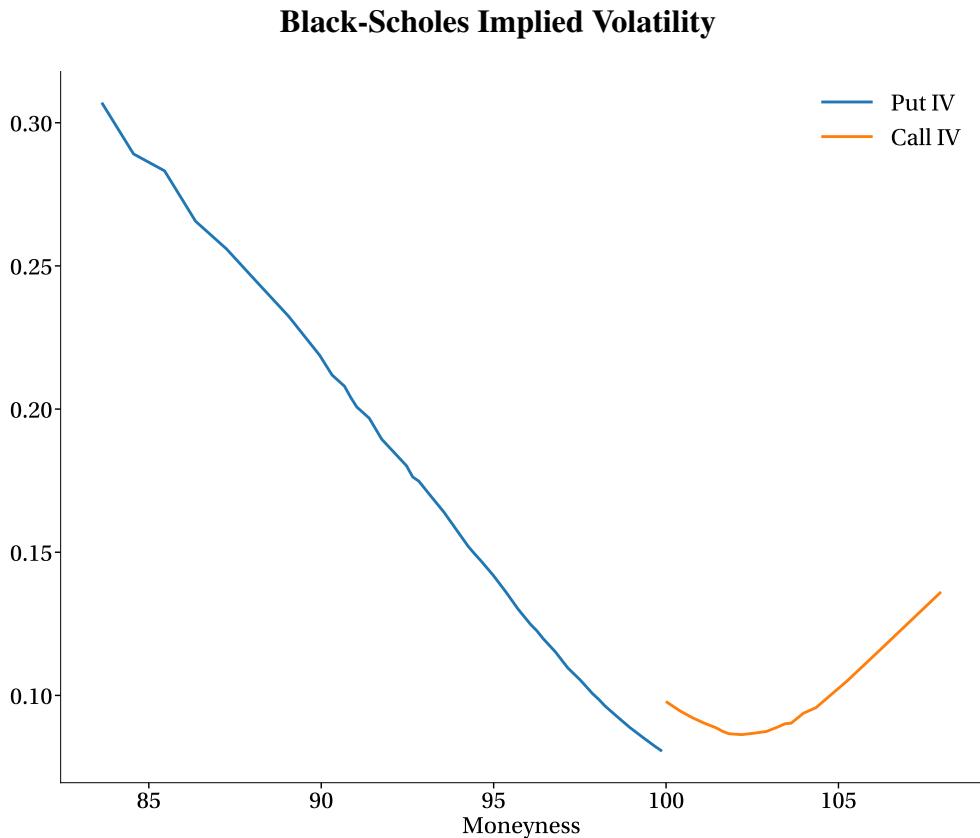


Figure 7.12: Plot of the Black-Scholes implied volatility "smile" on January 15, 2018, based on options on SPY expiring on February 2, 2018.

- Leverage effects are ruled out: Leverage, or negative correlation between the price and volatility of an asset, can generate negative skewness. This feature of asset prices increases the probability of extreme negative returns relative to the log-normal price process assumed in the B-S option pricing formula.
- No jumps: Jumps are also an empirical fact of most asset prices. Jumps, like time-varying volatility, increase the chance of seeing an extreme return.

The consequences of these limits are that, contrary to what the model underlying the B-S implies, B-S implied volatilities are not constant across strike prices, and so cannot be interpreted as market-based estimates of volatility.

Model-free implied volatility (MFIV) was developed as an alternative to B-S implied volatility by Demeterfi et al. (1999) and Britten-Jones and Neuberger (2000) with an important extension to jump processes and practical implementation details provided by Jiang and Tian (2005). These estimators build on Breeden and Litzenberger (1978) which contains key result that demonstrates how option prices are related to the risk-neutral measure – the distribution of asset price returns after removing risk premia. Suppose that the risk-neutral measure \mathbb{Q} exists and is unique. Then, under the risk-neutral measure, it must be the case that

$$\frac{\partial S_t}{S_t} = \sigma(t, \cdot) dW_t \quad (7.119)$$

is a martingale where $\sigma(t, \cdot)$ is a (possibly) time-varying volatility process that may depend on the stock price or other state variables. From the relationship, the price of a call option can be computed as

$$C(t, K) = E_{\mathbb{Q}} \left[(S_t - K)^+ \right] \quad (7.120)$$

for $t > 0, K > 0$ where the function $(x)^+ = \max(x, 0)$. Thus

$$C(t, K) = \int_K^{\infty} (S_t - K) \phi_t(S_t) dS_t \quad (7.121)$$

where $\phi_t(\cdot)$ is the risk-neutral measure. Differentiating with respect to K ,

$$\frac{\partial C(t, K)}{\partial K} = - \int_K^{\infty} \phi_t(S_t) dS_t. \quad (7.122)$$

Differentiating this expression again with respect to K (note K in the lower integral bound),

$$\frac{\partial^2 C(t, K)}{\partial K^2} = \phi_t(K), \quad (7.123)$$

and so that the risk-neutral density can be recovered from options prices. This result provides a basis for nonparametrically estimating the risk-neutral density from observed options prices (see, e.g., Aït-Sahalia and Lo (1998)). Another consequence of this result is that the expected (under \mathbb{Q}) variation in a stock price over the interval $[t_1, t_2]$ measure can be recovered from

$$E_{\mathbb{Q}} \left[\int_{t_1}^{t_2} \left(\frac{\partial S_t}{S_t} \right)^2 \right] = 2 \int_0^{\infty} \frac{C(t_2, K) - C(t_1, K)}{K^2} dK. \quad (7.124)$$

This expression cannot be directly implemented to recover the expected volatility since it requires a continuum of strike prices.

Equation 7.124 assumes that the risk-free rate is 0. When it is not, a similar result can be derived using the forward price

$$E_{\mathbb{F}} \left[\int_{t_1}^{t_2} \left(\frac{\partial F_t}{F_t} \right)^2 \right] = 2 \int_0^{\infty} \frac{C^F(t_2, K) - C^F(t_1, K)}{K^2} dK \quad (7.125)$$

where \mathbb{F} is the forward probability measure – that is, the probability measure where the forward price is a martingale and $C^F(\cdot, \cdot)$ is used to denote that this option is defined on the forward price. Additionally, when t_1 is 0, as is usually the case, the expression simplifies to

$$E_{\mathbb{F}} \left[\int_0^t \left(\frac{\partial F_t}{F_t} \right)^2 \right] = 2 \int_0^{\infty} \frac{C^F(t, K) - (F_0 - K)^+}{K^2} dK. \quad (7.126)$$

A number of important caveats are needed for employing this relationship to compute MFIV from option prices:

- Spot rather than forward prices. Because spot prices are usually used rather than forwards, the dependent variable needs to be redefined. If interest rates are non-stochastic, then define $B(0, T)$ to be the price of a bond today that pays \$1 time T . Thus, $F_0 = S_0/B(0, T)$, is the forward price and $C^F(T, K) = C(T, K)/B(0, T)$ is the forward option price. With the assumption of non-stochastic interest rates, the model-free implied volatility can be expressed

$$E_{\mathbb{F}} \left[\int_0^t \left(\frac{\partial S_t}{S_t} \right)^2 \right] = 2 \int_0^\infty \frac{C(t, K)/B(0, T) - (S_0/B(0, T) - K)^+}{K^2} dK \quad (7.127)$$

or equivalently using a change of variables as

$$E_{\mathbb{F}} \left[\int_0^t \left(\frac{\partial S_t}{S_t} \right)^2 \right] = 2 \int_0^\infty \frac{C(t, K/B(0, T)) - (S_0 - K)^+}{K^2} dK. \quad (7.128)$$

- Discretization. Because only finitely many options prices are available, the integral must be approximated using a discrete grid. Thus the approximation

$$E_{\mathbb{F}} \left[\int_0^t \left(\frac{\partial S_t}{S_t} \right)^2 \right] = 2 \int_0^\infty \frac{C(t, K/B(0, T)) - (S_0 - K)^+}{K^2} dK \quad (7.129)$$

$$\approx \sum_{m=1}^M [g(T, K_m) + g(T, K_{m-1})] (K_m - K_{m-1}) \quad (7.130)$$

The is used where

$$g(T, K) = \frac{C(t, K/B(0, T)) - (S_0 - K)^+}{K^2} \quad (7.131)$$

If the option tree is rich, this should not pose a significant issue. For option trees on individual firms, asset-specific study (for example, using data-calibrated Monte Carlo experiment) may be needed to ascertain whether the MFIV is a good estimate of the volatility under the forward measure.

- Maximum and minimum strike prices. The integral cannot be implemented from 0 to ∞ , and so the implied volatility has a downward bias due to the effect of the tails. In rich options trees, such as for the S&P 500, this issue is minor.

7.9.3 VIX

The VIX – Volatility Index – is a volatility measure produced by the Chicago Board Options Exchange (CBOE). It is computed using a “model-free” like estimator which uses both call and put prices.²² The VIX is an estimator of the price of a variance swap, which applies put-call parity to the previous expression to produce

$$\frac{2}{T} \exp(rT) \left(\int_0^{F_0} \frac{P(t, K/B(0, T))}{K^2} dK + \int_{F_0}^\infty \frac{C(t, K/B(0, T))}{K^2} dK \right).$$

²²The VIX is based exclusively on out-of-the-money prices, so calls are used for strikes above the current price and puts are used for strikes below the current price.

The term $(S_0 - K)^+$ drops out of this expression since it only used out-of-the-money options.

The VIX is computed according to

$$\sigma^2 = \frac{2}{T} \exp(rT) \sum_{i=1}^N \frac{\Delta K_i}{K_i^2} Q(K_i) - \frac{1}{T} \left(\frac{F_0}{K_0} - 1 \right)^2 \quad (7.132)$$

where T is the time to expiration of the options used, F_0 is the forward price which is computed from index option prices, K_i is the strike of the i^{th} out-of-the-money option, $\Delta K_i = (K_{i+1} - K_{i-1})/2$ is half of the distance of the interval surrounding the option with a strike price of K_i , K_0 is the strike of the option immediately below the forward level, F_0 , r is the risk-free rate and $Q(K_i)$ is the mid-point of the bid and ask for the call or put used at strike K_i . The forward index price is extracted using put-call parity as $F_0 = K_0 + \exp(rT)(C_0 - P_0)$ where K_0 is the strike price where the price difference between put and call is smallest, and C_0 and P_0 are, respectively, the call and put prices at this node. The VIX is typically calculated from options at the two maturities closes to the 30-day horizon (for example 28- and 35-days when using options that expire weekly). More details on the implementation of the VIX can be found in the CBOE whitepaper (CBOE, 2003).

The first term in the formula for the VIX can be viewed as

$$\frac{\Delta K_i}{K_i^2} Q(K_i) = \underbrace{\frac{\Delta K_i}{K_i}}_{\% \text{ width of interval}} \times \underbrace{\frac{Q(K_i)}{K_i}}_{\% \text{ option premium}},$$

so that the implied variance depends on only the option premium as a percent of the strike price. The division in the second term by K_0 similarly transforms the forward price to a percentage of strike measure. Each of these terms is width time height (premium), and so the VIX is the area below the out-of-the-money option pricing curve. When volatility is higher, all options are more valuable, and so there is more area below the curve. Figure 7.13 illustrates this area using option prices computed from the Black-Scholes formula for volatilities of 20% and 60%.

7.9.4 Computing the VIX from Black-Scholes prices

Put and call options values were computed from the Black-Scholes option pricing formula for an underlying with a price of \$100, an option time to maturity of a month ($T = 1/12$), a volatility of 20%, and a risk-free rate of 2%. Figure 7.13 plots the put and call options values from the Black-Scholes formula. The solid lines indicate the options that are out-of-the-money – puts with strike prices below \$100 or calls with strikes above \$100 – that are used to compute the VIX. The dotted lines show the option prices that are in-the-money. The values in Table 7.7 show all strikes where the out-of-the-month option price was at least \$0.01. These values are marked in Figure 7.13. The VIX is computed using the out-of-the-money option price $Q(K_i)$ rescaled by $2/T \exp(rT) \Delta K_i / K_i^2 = 2/1/12 \exp(.02/12) \times 4/K_i$ since the strikes are measured every \$4. The final line shows the total – 0.0430. The VIX index computed from these values is then $100 \times \sqrt{0.0430 - 3.338 \times 10^{-5}}\% = 20.75\%$, which is close to the true value of 20%. The second term in the square root is the adjustment $1/T (F/K_0 - 1)^2$ which is small. The small difference between the MFIV and the true volatility of 20% is due to discretization error since the strikes are only observed every \$4 and truncation error since only options with values larger than \$0.01 were used. The bottom panel of Figure 7.13 plots the option prices and highlights the area estimated by the VIX formula when the asset price volatility is 60%.

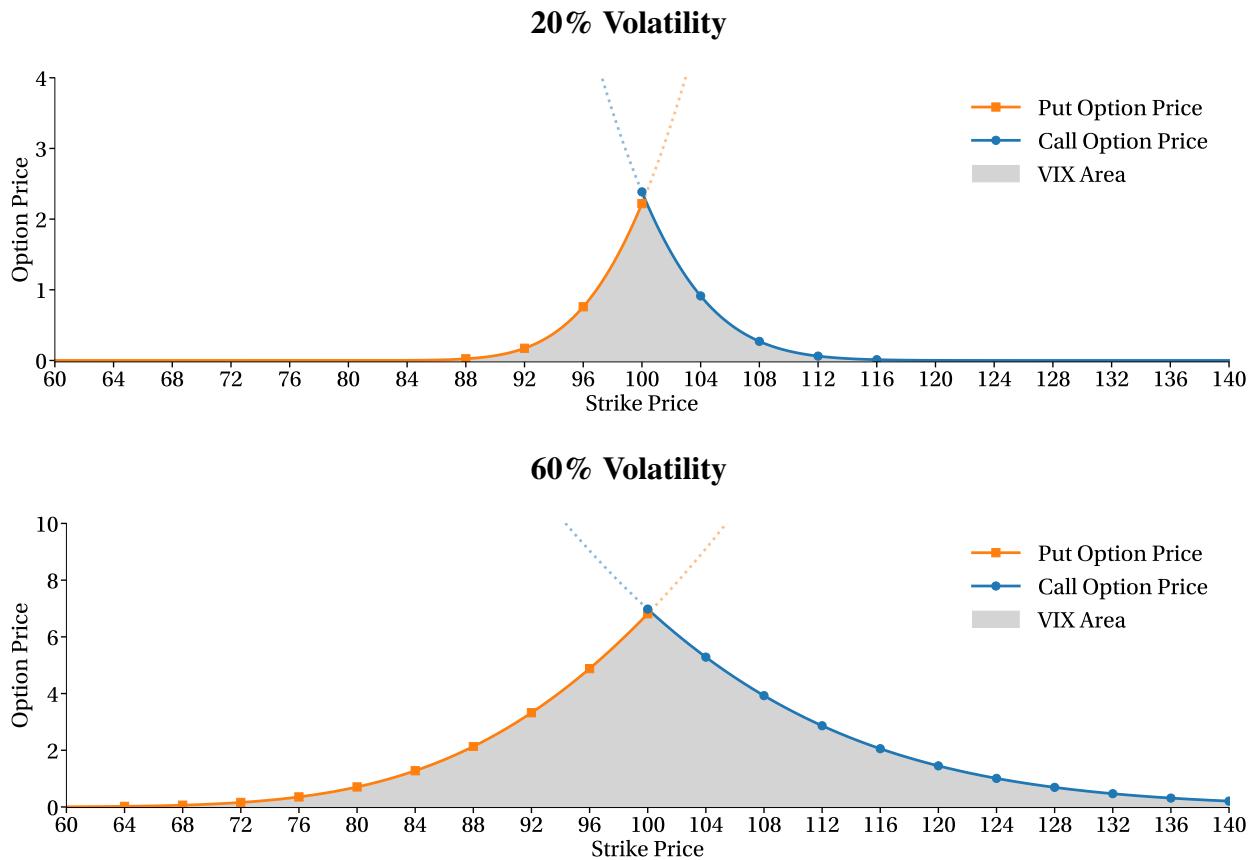


Figure 7.13: Option prices generated from the Black-Scholes pricing formula for an underlying with a price of \$100 with a volatility of 20% or 60% (bottom). The options expire in 1 month ($T = 1/12$), and the risk-free rate is 2%. The solid lines show the out-of-the-money options that are used to compute the VIX. The solid markers show the values where the option price to be at least \$0.01 using a \$4 grid of strike prices.

Strike	Call	Put	Abs. Diff.	VIX Contrib.
88	12.17	0.02	12.15	0.0002483
92	8.33	0.17	8.15	0.0019314
96	4.92	0.76	4.16	0.0079299
100	2.39	2.22	0.17	0.0221168
104	0.91	4.74	3.83	0.0080904
108	0.27	8.09	7.82	0.0022259
112	0.06	11.88	11.81	0.0004599
116	0.01	15.82	15.81	7.146e-05
Total				0.0430742

Table 7.7: Option prices generated from the Black-Scholes pricing formula for an underlying with a price of \$100 with a volatility of 20%. The options expire in 1 month ($T = 1/12$), and the risk-free rate is 2%. The third column shows the absolute difference which is used to determine K_0 in the VIX formula. The final column contains the contribution of each option to the VIX as measured by $2/T \exp(rT) \Delta K_i / K_i^2 \times Q(K_i)$.

7.9.5 Empirical Relationships

The daily VIX series from January 1990 until December 2018 is plotted in Figure 7.14 against a 22-day *forward* moving average computed as

$$\sigma_t^{MA} = \sqrt{\frac{252}{22} \sum_{i=0}^{21} r_{t+i}^2}.$$

The second panel shows the difference between these two series. The VIX is consistently, but not uniformly, higher than the forward volatility. This relationship highlights both a feature and a drawback of using a measure of the volatility computed under the risk-neutral measure: it captures a (possibly) time-varying risk premium. This risk premium captures investor compensation for changes in volatility (volatility of volatility) and jump risks.

7.A Kurtosis of an ARCH(1)

The necessary steps to derive the kurtosis of an ARCH(1) process are

$$\begin{aligned}
E[\epsilon_t^4] &= E[E_{t-1}[\epsilon_t^4]] \\
&= E[3(\omega + \alpha_1 \epsilon_{t-1}^2)^2] \\
&= 3E[(\omega + \alpha_1 \epsilon_{t-1}^2)^2] \\
&= 3E[\omega^2 + 2\omega\alpha_1 \epsilon_{t-1}^2 + \alpha_1^2 \epsilon_{t-1}^4] \\
&= 3(\omega^2 + \omega\alpha_1 E[\epsilon_{t-1}^2] + \alpha_1^2 E[\epsilon_{t-1}^4]) \\
&= 3\omega^2 + 6\omega\alpha_1 E[\epsilon_{t-1}^2] + 3\alpha_1^2 E[\epsilon_{t-1}^4].
\end{aligned} \tag{7.133}$$

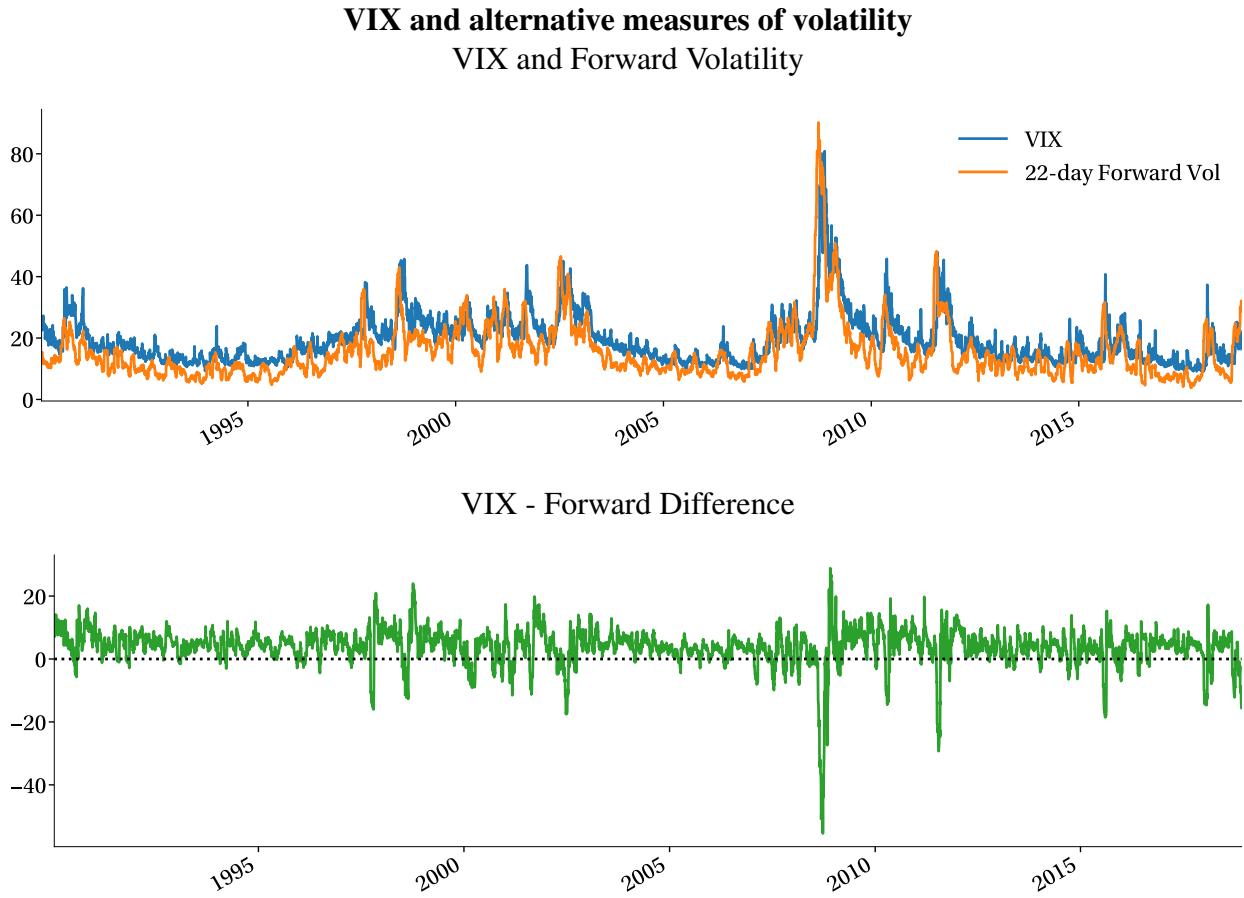


Figure 7.14: Plots of the VIX against a TARCH-based estimate of the volatility (top panel) and a 22-day forward moving average (bottom panel). The VIX is consistently above both measures reflecting the presence of a risk premium that compensates for time-varying volatility and jumps in the market return.

Using μ_4 to represent the expectation of the fourth power of ε_t ($\mu_4 = E[\varepsilon_t^4]$),

$$\begin{aligned}
 E[\varepsilon_t^4] - 3\alpha_1^2 E[\varepsilon_{t-1}^4] &= 3\omega^2 + 6\omega\alpha_1 E[\varepsilon_{t-1}^2] \\
 \mu_4 - 3\alpha_1^2 \mu_4 &= 3\omega^2 + 6\omega\alpha_1 \bar{\sigma}^2 \\
 \mu_4(1 - 3\alpha_1^2) &= 3\omega^2 + 6\omega\alpha_1 \bar{\sigma}^2 \\
 \mu_4 &= \frac{3\omega^2 + 6\omega\alpha_1 \bar{\sigma}^2}{1 - 3\alpha_1^2} \\
 \mu_4 &= \frac{3\omega^2 + 6\omega\alpha_1 \frac{\omega}{1-\alpha_1}}{1 - 3\alpha_1^2} \\
 \mu_4 &= \frac{3\omega^2(1 + 2\frac{\alpha_1}{1-\alpha_1})}{1 - 3\alpha_1^2}
 \end{aligned} \tag{7.134}$$

$$\mu_4 = \frac{3\omega^2(1+\alpha_1)}{(1-3\alpha_1^2)(1-\alpha_1)}.$$

This derivation makes use of the same principals as the intuitive proof and the identity that $\bar{\sigma}^2 = \omega/(1-\alpha_1)$. The final form highlights two important issues: first, μ_4 (and thus the kurtosis) is only finite if $1-3\alpha_1^2 > 0$ which requires that $\alpha_1 < \sqrt{\frac{1}{3}} \approx .577$, and second, the kurtosis, $\kappa = \frac{E[\varepsilon_t^4]}{E[\varepsilon_t^2]^2} = \frac{\mu_4}{\bar{\sigma}^2}$, is always greater than 3 since

$$\begin{aligned}\kappa &= \frac{E[\varepsilon_t^4]}{E[\varepsilon_t^2]^2} \\ &= \frac{\frac{3\omega^2(1+\alpha_1)}{(1-3\alpha_1^2)(1-\alpha_1)}}{\frac{\omega^2}{(1-\alpha_1)^2}} \\ &= \frac{3(1-\alpha_1)(1+\alpha_1)}{(1-3\alpha_1^2)} \\ &= \frac{3(1-\alpha_1^2)}{(1-3\alpha_1^2)} > 3.\end{aligned}\tag{7.135}$$

Finally, the variance of ε_t^2 can be computed noting that for any variable Y , $V[Y] = E[Y^2] - E[Y]^2$, and so

$$\begin{aligned}V[\varepsilon_t^2] &= E[\varepsilon_t^4] - E[\varepsilon_t^2]^2 \\ &= \frac{3\omega^2(1+\alpha_1)}{(1-3\alpha_1^2)(1-\alpha_1)} - \frac{\omega^2}{(1-\alpha_1)^2} \\ &= \frac{3\omega^2(1+\alpha_1)(1-\alpha_1)^2}{(1-3\alpha_1^2)(1-\alpha_1)(1-\alpha_1)^2} - \frac{\omega^2(1-3\alpha_1^2)(1-\alpha_1)}{(1-3\alpha_1^2)(1-\alpha_1)(1-\alpha_1)^2} \\ &= \frac{3\omega^2(1+\alpha_1)(1-\alpha_1)^2 - \omega^2(1-3\alpha_1^2)(1-\alpha_1)}{(1-3\alpha_1^2)(1-\alpha_1)(1-\alpha_1)^2} \\ &= \frac{3\omega^2(1+\alpha_1)(1-\alpha_1) - \omega^2(1-3\alpha_1^2)}{(1-3\alpha_1^2)(1-\alpha_1)^2} \\ &= \frac{3\omega^2(1-\alpha_1^2) - \omega^2(1-3\alpha_1^2)}{(1-3\alpha_1^2)(1-\alpha_1)^2} \\ &= \frac{3\omega^2(1-\alpha_1^2) - 3\omega^2(\frac{1}{3}-\alpha_1^2)}{(1-3\alpha_1^2)(1-\alpha_1)^2} \\ &= \frac{3\omega^2[(1-\alpha_1^2) - (\frac{1}{3}-\alpha_1^2)]}{(1-3\alpha_1^2)(1-\alpha_1)^2} \\ &= \frac{2\omega^2}{(1-3\alpha_1^2)(1-\alpha_1)^2} \\ &= \left(\frac{\omega}{1-\alpha_1}\right)^2 \frac{2}{(1-3\alpha_1^2)}\end{aligned}\tag{7.136}$$

$$= \frac{2\bar{\sigma}^4}{(1 - 3\alpha_1^2)}$$

The variance of the squared returns depends on the unconditional level of the variance, $\bar{\sigma}^2$, and the innovation term (α_1) squared.

7.B Kurtosis of a GARCH(1,1)

First, note that $E[\sigma_t^2 - \varepsilon_t^2] = 0$, so that $V[\sigma_t^2 - \varepsilon_t^2] = E[(\sigma_t^2 - \varepsilon_t^2)^2]$. This term can be expanded to $E[\varepsilon_t^4] - 2E[\varepsilon_t^2 \sigma_t^2] + E[\sigma_t^4]$ which can be shown to be $2E[\sigma_t^4]$ since

$$\begin{aligned} E[\varepsilon_t^4] &= E[E_{t-1}[e_t^4 \sigma_t^4]] \\ &= E[E_{t-1}[e_t^4] \sigma_t^4] \\ &= E[3\sigma_t^4] \\ &= 3E[\sigma_t^4] \end{aligned} \tag{7.137}$$

and

$$\begin{aligned} E[\varepsilon_t^2 \sigma_t^2] &= E[E_{t-1}[e_t^2 \sigma_t^2] \sigma_t^2] \\ &= E[\sigma_t^2 \sigma_t^2] \\ &= E[\sigma_t^4] \end{aligned} \tag{7.138}$$

so

$$\begin{aligned} E[\varepsilon_t^4] - 2E[\varepsilon_t^2 \sigma_t^2] + E[\sigma_t^4] &= 3E[\sigma_t^4] - 2E[\sigma_t^4] + E[\sigma_t^4] \\ &= 2E[\sigma_t^4] \end{aligned} \tag{7.139}$$

The only remaining step is to complete the tedious derivation of the expectation of this fourth power,

$$\begin{aligned} E[\sigma_t^4] &= E[(\sigma_t^2)^2] \\ &= E[(\omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2)^2] \\ &= E[\omega^2 + 2\omega\alpha_1 \varepsilon_{t-1}^2 + 2\omega\beta_1 \sigma_{t-1}^2 + 2\alpha_1\beta_1 \varepsilon_{t-1}^2 \sigma_{t-1}^2 + \alpha_1^2 \varepsilon_{t-1}^4 + \beta_1^2 \sigma_{t-1}^4] \\ &= \omega^2 + 2\omega\alpha_1 E[\varepsilon_{t-1}^2] + 2\omega\beta_1 E[\sigma_{t-1}^2] + 2\alpha_1\beta_1 E[\varepsilon_{t-1}^2 \sigma_{t-1}^2] + \alpha_1^2 E[\varepsilon_{t-1}^4] + \beta_1^2 E[\sigma_{t-1}^4] \end{aligned} \tag{7.140}$$

Noting that

- $E[\varepsilon_{t-1}^2] = E[E_{t-2}[\varepsilon_{t-1}^2]] = E[E_{t-2}[e_{t-1}^2 \sigma_{t-1}^2]] = E[\sigma_{t-1}^2 E_{t-2}[e_{t-1}^2]] = E[\sigma_{t-1}^2] = \bar{\sigma}^2$
- $E[\varepsilon_{t-1}^2 \sigma_{t-1}^2] = E[E_{t-2}[\varepsilon_{t-1}^2 \sigma_{t-1}^2]] = E[E_{t-2}[e_{t-1}^2 \sigma_{t-1}^2] \sigma_{t-1}^2] = E[E_{t-2}[e_{t-1}^2] \sigma_{t-1}^2 \sigma_{t-1}^2] = E[\sigma_{t-1}^4]$

- $E[\varepsilon_{t-1}^4] = E[E_{t-2}[\varepsilon_{t-1}^4]] = E[E_{t-2}[e_{t-1}^4 \sigma_{t-1}^4]] = 3E[\sigma_{t-1}^4]$

the final expression for $E[\sigma_t^4]$ can be arrived at

$$\begin{aligned} E[\sigma_t^4] &= \omega^2 + 2\omega\alpha_1 E[\varepsilon_{t-1}^2] + 2\omega\beta_1 E[\sigma_{t-1}^2] + 2\alpha_1\beta_1 E[\varepsilon_{t-1}^2 \sigma_{t-1}^2] + \alpha_1^2 E[\varepsilon_{t-1}^4] + \beta_1^2 E[\sigma_{t-1}^4] \quad (7.141) \\ &= \omega^2 + 2\omega\alpha_1 \bar{\sigma}^2 + 2\omega\beta_1 \bar{\sigma}^2 + 2\alpha_1\beta_1 E[\sigma_{t-1}^4] + 3\alpha_1^2 E[\sigma_{t-1}^4] + \beta_1^2 E[\sigma_{t-1}^4]. \end{aligned}$$

$E[\sigma_t^4]$ can be solved for (replacing $E[\sigma_t^4]$ with μ_4),

$$\begin{aligned} \mu_4 &= \omega^2 + 2\omega\alpha_1 \bar{\sigma}^2 + 2\omega\beta_1 \bar{\sigma}^2 + 2\alpha_1\beta_1 \mu_4 + 3\alpha_1^2 \mu_4 + \beta_1^2 \mu_4 \quad (7.142) \\ \mu_4 - 2\alpha_1\beta_1 \mu_4 - 3\alpha_1^2 \mu_4 - \beta_1^2 \mu_4 &= \omega^2 + 2\omega\alpha_1 \bar{\sigma}^2 + 2\omega\beta_1 \bar{\sigma}^2 \\ \mu_4(1 - 2\alpha_1\beta_1 - 3\alpha_1^2 - \beta_1^2) &= \omega^2 + 2\omega\alpha_1 \bar{\sigma}^2 + 2\omega\beta_1 \bar{\sigma}^2 \\ \mu_4 &= \frac{\omega^2 + 2\omega\alpha_1 \bar{\sigma}^2 + 2\omega\beta_1 \bar{\sigma}^2}{1 - 2\alpha_1\beta_1 - 3\alpha_1^2 - \beta_1^2} \end{aligned}$$

finally substituting $\bar{\sigma}^2 = \omega/(1 - \alpha_1 - \beta_1)$ and returning to the original derivation,

$$E[\varepsilon_t^4] = \frac{3(1 + \alpha_1 + \beta_1)}{(1 - \alpha_1 - \beta_1)(1 - 2\alpha_1\beta_1 - 3\alpha_1^2 - \beta_1^2)}, \quad (7.143)$$

and the kurtosis, $\kappa = \frac{E[\varepsilon_t^4]}{E[\varepsilon_t^2]^2} = \frac{\mu_4}{\bar{\sigma}^2}$, which simplifies to

$$\kappa = \frac{3(1 + \alpha_1 + \beta_1)(1 - \alpha_1 - \beta_1)}{1 - 2\alpha_1\beta_1 - 3\alpha_1^2 - \beta_1^2} > 3. \quad (7.144)$$

Short Problems

Problem 7.1. What is Realized Variance and why is it useful?

Problem 7.2. Suppose $r_t = \sigma_t \varepsilon_t$ where $\sigma_t^2 = \omega + \alpha r_{t-1}^2 + \beta \sigma_{t-1}^2$, and $\varepsilon_t \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$. What conditions are required on the parameters ω , α , and β for r_t to be covariance stationary?

Problem 7.3. What is Realized Variance?

Problem 7.4. Discuss the properties of the generalized forecast error from a correctly specified volatility model.

Problem 7.5. Outline the steps the in Mincer-Zarnowitz framework to objectively evaluate a sequence of variance forecasts $\{\hat{\sigma}_{t+1|t}^2\}$.

Exercises

Exercise 7.1. Suppose we model log-prices at time t , written p_t , as an ARCH(1) process

$$p_t | \mathcal{F}_{t-1} \sim N(p_{t-1}, \sigma_t^2),$$

where \mathcal{F}_t denotes the information up to and including time t and

$$\sigma_t^2 = \alpha + \beta (p_{t-1} - p_{t-2})^2.$$

1. Is p_t a martingale?

2. What is

$$E[\sigma_t^2]?$$

3. Calculate

$$\text{Cov} \left[(p_t - p_{t-1})^2, (p_{t-s} - p_{t-1-s})^2 \right]$$

for $s > 0$.

4. Comment on the importance of this result from a practical perspective.

5. How do you use a likelihood function to estimate an ARCH model?

6. How can the ARCH(1) model be generalized better capture the variance dynamics of asset prices?

7. In the ARCH(1) case, what can you say about the properties of

$$p_{t+s} | \mathcal{F}_{t-1},$$

for $s > 0$, i.e., the multi-step ahead forecast of prices?

8. Why are Bollerslev-Wooldridge standard errors important when testing coefficients in ARCH models?

Exercise 7.2. Derive explicit relationships between the parameters of an APARCH(1,1,1),

$$\begin{aligned} r_t &= \mu_t + \varepsilon_t \\ \sigma_t^\delta &= \omega + \alpha_1 (|\varepsilon_{t-1}| + \gamma_1 \varepsilon_{t-1})^\delta + \beta_1 \sigma_{t-1}^\delta \\ \varepsilon_t &= \sigma_t e_t \\ e_t &\stackrel{\text{i.i.d.}}{\sim} N(0, 1), \end{aligned}$$

and:

1. ARCH(1)
2. GARCH(1,1)
3. AVGARCH(1,1)

4. TARCH(1,1,1)
5. GJR-GARCH(1,1,1)

Exercise 7.3. Consider the following GJR-GARCH process,

$$\begin{aligned} r_t &= \rho r_{t-1} + \varepsilon_t \\ \varepsilon_t &= \sigma_t e_t \\ \sigma_t^2 &= \omega + \alpha \varepsilon_{t-1}^2 + \gamma \varepsilon_{t-1}^2 I_{[\varepsilon_{t-1} < 0]} + \beta \sigma_{t-1}^2 \\ e_t &\stackrel{\text{i.i.d.}}{\sim} N(0, 1) \end{aligned}$$

where $E_t[\cdot] = E[\cdot | \mathcal{F}_t]$ is the time t conditional expectation and $V_t[\cdot] = V[\cdot | \mathcal{F}_t]$ is the time t conditional variance.

1. What conditions are necessary for this process to be covariance stationary?

Assume these conditions hold in the remaining questions. *Note:* If you cannot answer one or more of these questions for an arbitrary γ , you can assume that $\gamma = 0$ and receive partial credit.

2. What is $E[r_{t+1}]$?
3. What is $E_t[r_{t+1}]$?
4. What is $V[r_{t+1}]$?
5. What is $V_t[r_{t+1}]$?
6. What is $V_t[r_{t+2}]$?

Exercise 7.4. Let r_t follow a GARCH process

$$\begin{aligned} r_t &= \sigma_t e_t \\ \sigma_t^2 &= \omega + \alpha r_{t-1}^2 + \beta \sigma_{t-1}^2 \\ e_t &\stackrel{\text{i.i.d.}}{\sim} N(0, 1) \end{aligned}$$

1. What are the values of the following quantities?
 - (a) $E[r_{t+1}]$
 - (b) $E_t[r_{t+1}]$
 - (c) $V[r_{t+1}]$
 - (d) $V_t[r_{t+1}]$
 - (e) $\rho_1 = \text{Corr}[r_t, r_{t-1}]$
2. What is $E[(r_t^2 - \bar{\sigma}^2)(r_{t-1}^2 - \bar{\sigma}^2)]$ where $\bar{\sigma} = E[\sigma_t^2]$. Hint: Consider the relationship to ARMA models.
3. Describe the h -step ahead forecast from this model.

Exercise 7.5. Let r_t follow an ARCH process

$$\begin{aligned} r_t &= \sigma_t e_t \\ \sigma_t^2 &= \omega + \alpha_1 r_{t-1}^2 \\ e_t &\stackrel{\text{i.i.d.}}{\sim} N(0, 1) \end{aligned}$$

1. What are the values of the following quantities?
 - (a) $E[r_{t+1}]$
 - (b) $E_t[r_{t+1}]$
 - (c) $V[r_{t+1}]$
 - (d) $V_t[r_{t+1}]$
 - (e) $\rho_1 = \text{Corr}[r_t, r_{t-1}]$
2. What is $E[(r_t^2 - \bar{\sigma}^2)(r_{t-1}^2 - \bar{\sigma}^2)]$ where $\bar{\sigma} = E[\sigma_t^2]$. Hint: Think about the AR duality.
3. Describe the h -step ahead forecast from this model.

Exercise 7.6. Consider an EGARCH(1,1,1) model:

$$\ln \sigma_t^2 = \omega + \alpha_1 \left(|e_{t-1}| - \sqrt{\frac{2}{\pi}} \right) + \gamma_1 e_{t-1} + \beta_1 \ln \sigma_{t-1}^2$$

where $e_t \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$.

1. What are the required conditions on the model parameters for this process to be covariance stationary?
2. What is the one-step-ahead forecast of σ_t^2 , $E_t [\sigma_{t+1}^2]$?
3. What is the most you can say about the two-step-ahead forecast of σ_t^2 ($E_t [\sigma_{t+2}^2]$)?

Exercise 7.7. Answer the following questions:

1. Describe three fundamentally different procedures to estimate the volatility over some interval. What are the strengths and weaknesses of each?
2. Why is Realized Variance useful when evaluating a volatility model?
3. What considerations are important when computing Realized Variance?
4. Why does the Black-Scholes implied volatility vary across strikes?

Exercise 7.8. Consider a general volatility specification for an asset return r_t :

$$\begin{aligned} r_t | \mathcal{F}_{t-1} &\sim N(0, \sigma_t^2) \\ \text{and let } e_t &\equiv \frac{r_t}{\sigma_t} \\ \text{so } e_t | \mathcal{F}_{t-1} &\stackrel{\text{i.i.d.}}{\sim} N(0, 1) \end{aligned}$$

1. Find the conditional kurtosis of the returns:

$$\text{Kurt}_{t-1}[r_t] \equiv \frac{\text{E}_{t-1}[(r_t - \text{E}_{t-1}[r_t])^4]}{(\text{V}_{t-1}[r_t])^2}$$

2. Show that if $\text{V}[\sigma_t^2] > 0$, then the *unconditional* kurtosis of the returns,

$$\text{Kurt}[r_t] \equiv \frac{\text{E}[(r_t - \text{E}[r_t])^4]}{(\text{V}[r_t])^2}$$

is greater than 3.

3. Find the conditional skewness of the returns:

$$\text{Skew}_{t-1}[r_t] \equiv \frac{\text{E}_{t-1}[(r_t - \text{E}_{t-1}[r_t])^3]}{(\text{V}_{t-1}[r_t])^{3/2}}$$

4. Find the *unconditional* skewness of the returns:

$$\text{Skew}[r_t] \equiv \frac{\text{E}[(r_t - \text{E}[r_t])^3]}{(\text{V}[r_t])^{3/2}}$$

Exercise 7.9.

Answer the following questions:

1. Describe three fundamentally different procedures to estimate the volatility over some interval. What are the strengths and weaknesses of each?
2. Why does the Black-Scholes implied volatility vary across strikes?
3. Consider the following GJR-GARCH process,

$$\begin{aligned} r_t &= \mu + \rho r_{t-1} + \varepsilon_t \\ \varepsilon_t &= \sigma_t e_t \\ \sigma_t^2 &= \omega + \alpha \varepsilon_{t-1}^2 + \gamma \varepsilon_{t-1}^2 I_{[\varepsilon_{t-1} < 0]} + \beta \sigma_{t-1}^2 \\ e_t &\stackrel{\text{i.i.d.}}{\sim} N(0, 1) \end{aligned}$$

where $\text{E}_t[\cdot] = \text{E}[\cdot | \mathcal{F}_t]$ is the time t conditional expectation and $\text{V}_t[\cdot] = \text{V}[\cdot | \mathcal{F}_t]$ is the time t conditional variance.

- (a) What conditions are necessary for this process to be covariance stationary?

Assume these conditions hold in the remaining questions.

- (b) What is $\text{E}[r_{t+1}]$?
(c) What is $\text{E}_t[r_{t+1}]$?

- (d) What is $E_t[r_{t+2}]$?
- (e) What is $V[r_{t+1}]$?
- (f) What is $V_t[r_{t+1}]$?
- (g) What is $V_t[r_{t+2}]$?

Exercise 7.10. Answer the following questions about variance estimation.

1. What is Realized Variance?
2. How is Realized Variance estimated?
3. Describe two models which are appropriate for modeling Realized Variance.
4. What is an Exponential Weighted Moving Average (EWMA)?
5. Suppose an ARCH model for the conditional variance of daily returns was fit

$$\begin{aligned} r_{t+1} &= \mu + \sigma_{t+1} e_{t+1} \\ \sigma_{t+1}^2 &= \omega + \alpha_1 \varepsilon_t^2 + \alpha_2 \varepsilon_{t-1}^2 \\ e_t &\stackrel{\text{i.i.d.}}{\sim} N(0, 1) \end{aligned}$$

What are the forecasts for $t + 1$, $t + 2$ and $t + 3$ given the current (time t) information set?

6. Suppose an EWMA was used instead for the model of conditional variance with smoothing parameter = .94. What are the forecasts for $t + 1$, $t + 2$ and $t + 3$ given the current (time t) information set?
7. Compare the ARCH(2) and EWMA forecasts when the forecast horizon is large (e.g., $E_t[\sigma_{t+h}^2]$ for large h).
8. What is VIX?

Exercise 7.11. Suppose $\{y_t\}$ is covariance stationary and can be described by the following process:

$$\begin{aligned} y_t &= \phi_1 y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t \\ \varepsilon_t &= \sigma_t e_t \\ \sigma_t^2 &= \omega + \alpha \varepsilon_{t-1}^2 \\ e_t &\stackrel{\text{i.i.d.}}{\sim} N(0, 1) \end{aligned}$$

what are the values of the following quantities:

1. $E_t[y_{t+1}]$
2. $E_t[y_{t+2}]$
3. $\lim_{h \rightarrow \infty} E_t[y_{t+h}]$
4. $V_t[\varepsilon_{t+1}]$

5. $V_t [y_{t+1}]$
6. $V_t [y_{t+2}]$
7. $\lim_{h \rightarrow \infty} V_t [\varepsilon_{t+h}]$

Exercise 7.12. Answer the following questions:

Suppose $\{y_t\}$ is covariance stationary and can be described by the following process:

$$\begin{aligned} y_t &= \phi_0 + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \varepsilon_t \\ \varepsilon_t &= \sigma_t e_t \\ \sigma_t^2 &= \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \\ e_t &\stackrel{\text{i.i.d.}}{\sim} N(0, 1) \end{aligned}$$

what are the values of the following quantities:

1. $E_t [y_{t+1}]$
2. $E_t [y_{t+2}]$
3. $\lim_{h \rightarrow \infty} E_t [y_{t+h}]$
4. $V_t [\varepsilon_{t+1}]$
5. $V_t [y_{t+2}]$
6. $\lim_{h \rightarrow \infty} V_t [\varepsilon_{t+h}]$

Exercise 7.13. Consider the AR(2)-ARCH(2) model

$$\begin{aligned} y_t &= \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t \\ \varepsilon_t &= \sigma_t e_t \\ \sigma_t^2 &= \omega + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 \\ e_t &\stackrel{\text{i.i.d.}}{\sim} N(0, 1) \end{aligned}$$

1. What conditions are required for ϕ_0 , ϕ_1 and, ϕ_2 for the model to be covariance stationary?
2. What conditions are required for ω , α_1 , and α_2 for the model to be covariance stationary?
3. Show that $\{\varepsilon_t\}$ is a white noise process.
4. Are ε_t and ε_{t-s} independent for $s \neq 0$?
5. What are the values of the following quantities:
 - (a) $E[y_t]$
 - (b) $E_t [y_{t+1}]$
 - (c) $E_t [y_{t+2}]$

- (d) $V_t[y_{t+1}]$
 (e) $V_t[y_{t+2}]$

Exercise 7.14. Suppose $\{y_t\}$ is covariance stationary and can be described by the following process:

$$\begin{aligned} y_t &= \phi_1 y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t \\ \varepsilon_t &= \sigma_t e_t \\ \sigma_t^2 &= \omega + \alpha \varepsilon_{t-1}^2 \\ e_t &\stackrel{\text{i.i.d.}}{\sim} N(0, 1) \end{aligned}$$

1. What are the values of the following quantities:

- (a) $E_t[y_{t+1}]$
 (b) $E_t[y_{t+2}]$
 (c) $\lim_{h \rightarrow \infty} E_t[y_{t+h}]$
 (d) $V_t[\varepsilon_{t+1}]$
 (e) $V_t[y_{t+1}]$
 (f) $V_t[y_{t+2}]$
 (g) $V[y_{t+1}]$

Exercise 7.15. Consider the MA(2)-GARCH(1,1) model

$$\begin{aligned} y_t &= \phi_0 + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \varepsilon_t \\ \varepsilon_t &= \sigma_t e_t \\ \sigma_t^2 &= \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \\ e_t &\stackrel{\text{i.i.d.}}{\sim} N(0, 1) \end{aligned}$$

1. What conditions are required for ϕ_0 , θ_1 , and θ_2 for the model to be covariance stationary?
2. What conditions are required for ω , α_1 , and β_1 for the model to be covariance stationary?
3. Show that $\{\varepsilon_t\}$ is a white noise process.
4. Are ε_t and ε_{t-1} independent?
5. What are the values of the following quantities:

- (a) $E[y_t]$
 (b) $E_t[y_{t+1}]$
 (c) $E_t[y_{t+2}]$
 (d) $\lim_{h \rightarrow \infty} E_t[y_{t+h}]$
 (e) $V_t[y_{t+1}]$
 (f) $V_t[y_{t+2}]$

Exercise 7.16. Suppose $\{y_t\}$ is covariance stationary and can be described by the following process:

$$\begin{aligned} y_t &= \phi_0 + \phi_1 y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t \\ \varepsilon_t &= \sigma_t e_t \\ \sigma_t^2 &= \omega + \alpha \varepsilon_{t-1}^2 \\ e_t &\stackrel{\text{i.i.d.}}{\sim} N(0, 1) \end{aligned}$$

1. What are the values of the following quantities:

- (a) $E[y_{t+1}]$
- (b) $E_t[y_{t+1}]$
- (c) $E_t[y_{t+2}]$
- (d) $\lim_{h \rightarrow \infty} E_t[y_{t+h}]$
- (e) $V_t[\varepsilon_{t+1}]$
- (f) $V_t[y_{t+1}]$
- (g) $V_t[y_{t+2}]$
- (h) $V[y_{t+1}]$

2. Justify a reasonable model for each of these time series in Figure 7.15 using information in the autocorrelation and partial autocorrelation plots. In each set of plots, the leftmost panel shows that data ($T = 100$). The middle panel shows the sample autocorrelation with 95% confidence bands. The right panel shows the sample partial autocorrelation for the data with 95% confidence bands.

- (a) Panel (a)
- (b) Panel (b)

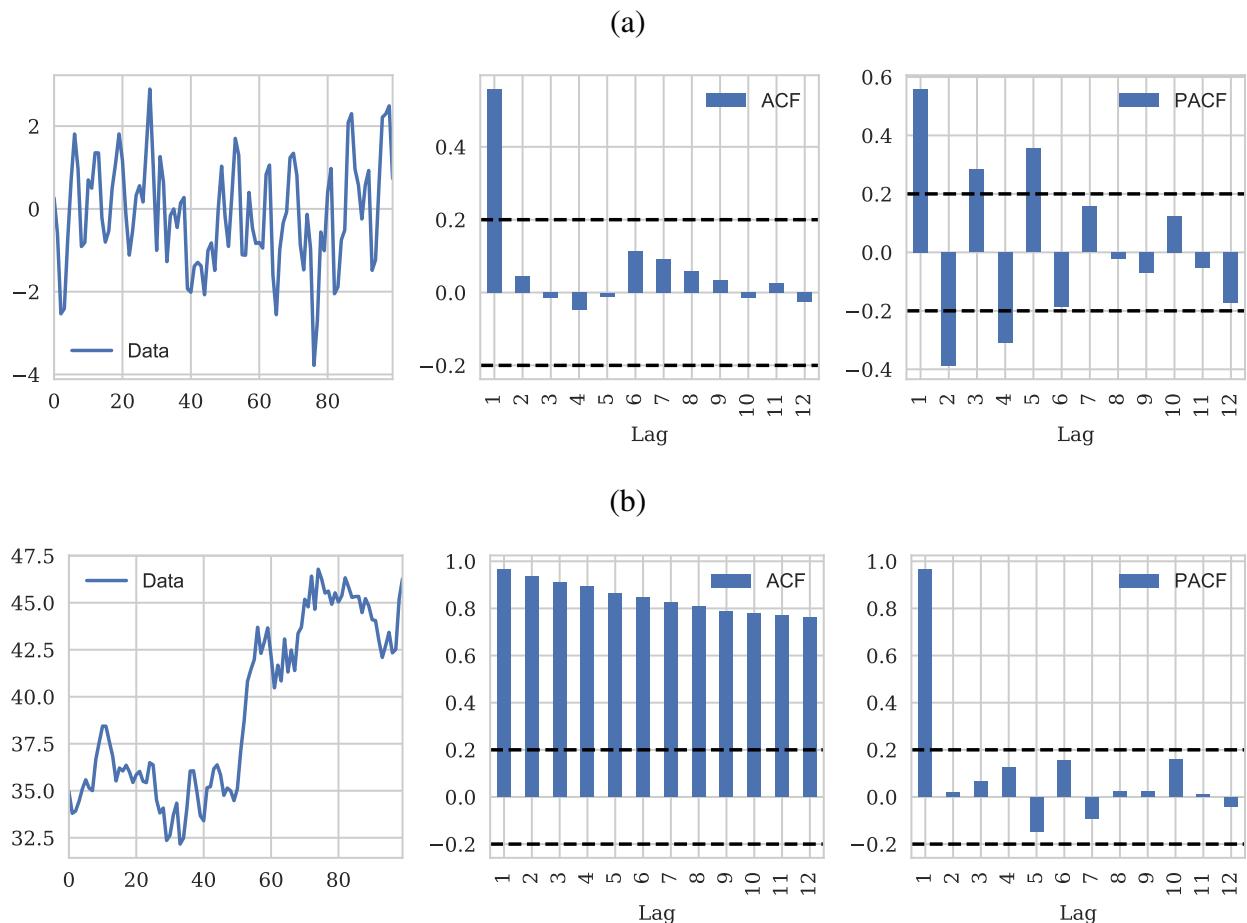


Figure 7.15: Plots for question 7.16.

Chapter 8

Value-at-Risk, Expected Shortfall and Density Forecasting

Alternative references for this chapter include Christoffersen (2003), which is a highly accessible introduction, Gourieroux and Jasiak (2009), who provide additional technical details, and, McNeil, Frey, and Embrechts (2005), who provide a comprehensive and technical treatment of risk measurement.

The American Heritage Dictionary, Fourth Edition, defines risk as “the possibility of suffering harm or loss; danger.” In finance, harm or loss has a specific meaning: decreases in the value of a portfolio. This chapter introduces three methods used to assess the riskiness of a portfolio: Value-at-Risk (VaR), Expected Shortfall, and modeling the entire density of the portfolio’s return.

8.1 Defining Risk

Portfolios are exposed to multiple distinct sources of risk. The most important sources of risk can be classified into one of six categories.

Market Risk

Market risk describes the uncertainty about the future price of an asset due to changes in fundamentals or beliefs. For example, market risk captures changes in asset prices due to macroeconomics announcements such as FOMC policy rate updates or non-farm payroll releases.

Liquidity risk

Liquidity risk complements market risk by measuring the loss involved if a position must be rapidly unwound. For example, if a fund wished to sell 20,000,000 shares of IBM on a single day, which has a typical daily volume of 5,000,000, this sale would be expected to have a substantial effect on the price. Liquidity risk is distinct from market risk since it represents a transitory distortion due to transaction pressure.

Credit Risk

Credit risk, also known as default risk, covers cases where a 3rd party is unable to pay per previously agreed to terms. Holders of corporate bonds are exposed to credit risk since the bond issuer may not be able to make some or all of the scheduled coupon payments.

Counterparty Risk

Counterparty risk extends credit risk to instruments other than bonds and captures the event that a counterparty to a transaction, for example, the seller of an option contract, is unable to complete the transaction at expiration. Counterparty risk was a significant factor in the financial crisis of 2008 where the protection offered in Credit Default Swaps (CDS) was not available when the underlying assets defaulted.

Model Risk

Model risk represents an econometric form of risk that measures the uncertainty about the correct form of the model used to compute the price of the asset or the asset's riskiness. Model risk is particularly important when prices of assets are primarily determined by a model rather than in a liquid market, as was the case in the Mortgage Backed Securities (MBS) market in 2007.

Estimation Risk

Estimation risk captures an aspect of risk that is present whenever estimated parameters are used in econometric models to price securities or assess risk. Estimation risk is distinct from model risk since it is present even if a model is correctly specified. In many practical applications, parameter estimation error can result in a substantial misstatement of risk. Model and estimation risk are always present and are generally substitutes – parsimonious models are more likely to be misspecified but may have less parameter estimation uncertainty.

This chapter deals exclusively with market risk. Liquidity, credit risk and counterparty risk all require specialized treatment beyond the scope of this course. Model evaluation, especially out-of-sample evaluation, is the primary tool for assessing model and estimation risks.

8.2 Value-at-Risk (VaR)

The most widely reported measure of risk is Value-at-Risk (VaR). The VaR of a portfolio is a measure of the risk in the left tail of portfolio's return over some period, often a day or a week. VaR provides a more sensible measure of the risk of the portfolio than variance since it focuses on losses. VaR is not a perfect measure of risk, and the issues with VaR are detailed in the context of coherent risk measures (section 8.8).

8.2.1 Value-at-Risk Defined

The VaR of a portfolio measures the value (in £, \$, €, ¥, ...) which an investor would lose with some small probability, usually between 1 and 10%, over a specified horizon. Because the VaR represents a potential *loss*, it is usually a positive number.

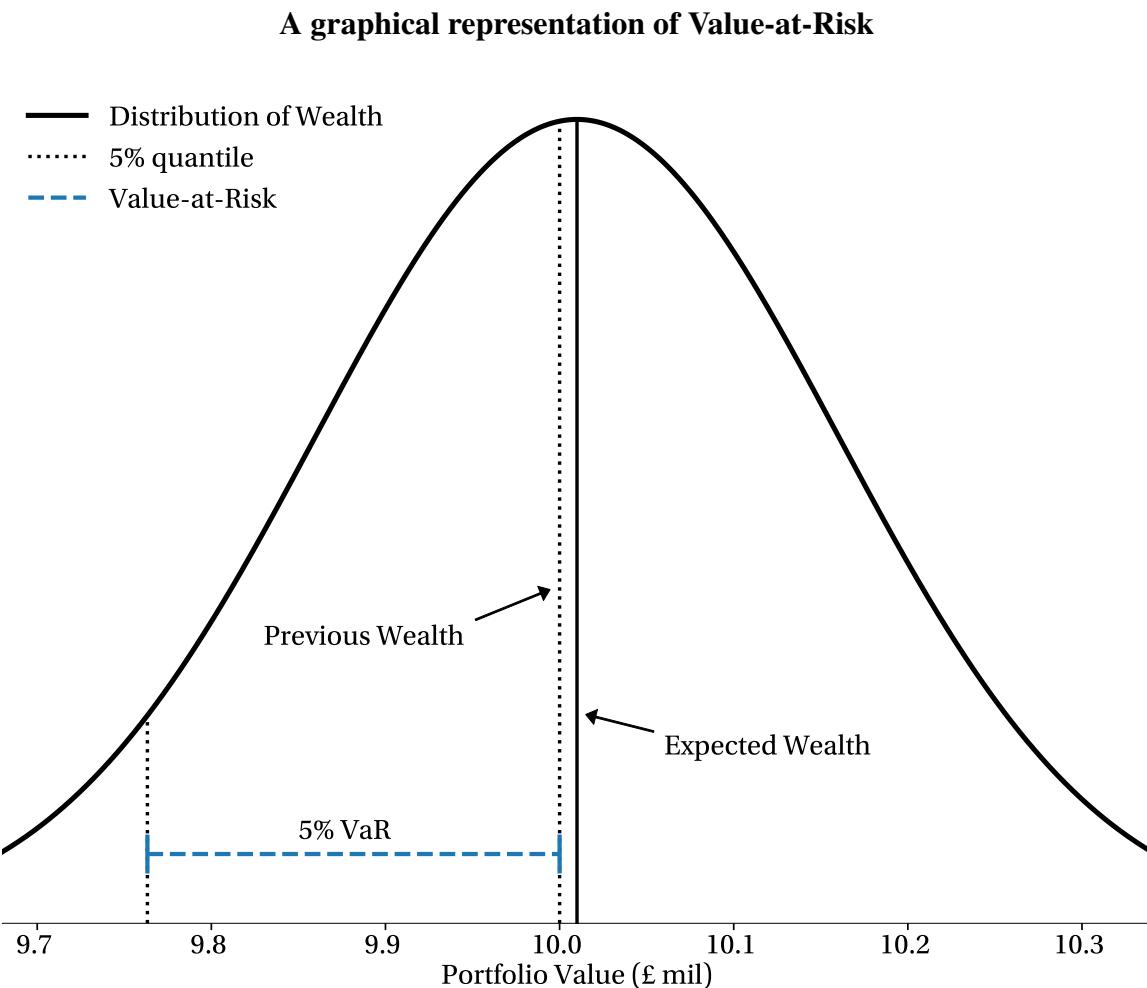


Figure 8.1: A graphical representation of Value-at-Risk. The VaR is represented by the magnitude of the horizontal bar and measures the distance between the value of the portfolio in the current period and the α -quantile of the portfolio value distribution. In this example, $\alpha = 5\%$, the value of the portfolio's assets is £10,000,000, and returns are $N(.001, .015^2)$.

Definition 8.1 (Value-at-Risk). The α Value-at-Risk (VaR) of a portfolio is defined as the largest change in the portfolio such that the probability that the loss in portfolio value over a specified horizon is greater than the VaR is α ,

$$\Pr(R_t < -\text{VaR}) = \alpha \quad (8.1)$$

where $R_t = W_t - W_{t-1}$ is the change in the value of the portfolio, W_t and the time span depends on the application (e.g., one day or two weeks).

For example, if an investor had a portfolio value of £10,000,000 and had a daily portfolio return which was $N(.001, .015^2)$ (annualized mean of 25%, volatility of 23.8%), the daily α Value-at-Risk of this portfolio is

$$\text{£}10,000,000(-.001 - .015\Phi^{-1}(\alpha)) = \text{£}236,728.04$$

where $\Phi^{-1}(\cdot)$ is the inverse CDF of a standard normal. This expression may appear backward – it is not. The negative sign on the mean indicates that increases in the mean *decrease* the VaR. The negative sign on the standard deviation term indicates that increases in the volatility raise the VaR since for $\alpha < .5$, $\Phi^{-1}(\alpha) < 0$. It is often more useful to express Value-at-Risk as a percentage of the portfolio value – e.g., 1.5% – rather than in units of currency since to remove the initial value portfolio from the measure.

Definition 8.2 (Percentage Value-at-Risk). The α percentage Value-at-Risk (%VaR) of a portfolio is defined as the largest return such that the probability that the return on the portfolio over a specified horizon is less than $-1 \times \% \text{VaR}$ is α ,

$$\Pr(r_t < -\% \text{VaR}) = \alpha \quad (8.2)$$

where $r_t = (W_t - W_{t-1}) / W_{t-1}$ is the return of the portfolio. %VaR can be equivalently defined as $\% \text{VaR} = \text{VaR} / W_{t-1}$.

Since percentage VaR and VaR only differ by the current value of the portfolio, the remainder of the chapter focuses on percentage VaR.

8.2.2 The relationship between VaR and quantiles

Understanding that VaR and quantiles are fundamentally related provides a key insight. If r is the return on a portfolio, the α -VaR is $-1 \times q_\alpha(r)$ where $q_\alpha(r)$ is the α -quantile of the portfolio's return. In most cases α is chosen to be some small quantile – 1, 5 or 10% – and so $q_\alpha(r)$ is a negative number.¹

8.3 Conditional Value-at-Risk

Most applications of VaR are used to measure risk over short horizons, and so require a conditional Value-at-Risk. Conditioning employs information up to time t to produce a VaR in period $t + h$.

Definition 8.3 (Conditional Value-at-Risk). The conditional α Value-at-Risk is defined as

$$\Pr(r_{t+1} < -\text{VaR}_{t+1|t} | \mathcal{F}_t) = \alpha \quad (8.3)$$

where $r_{t+1} = \frac{W_{t+1} - W_t}{W_t}$ is the time $t + 1$ return on a portfolio. Since t is an arbitrary measure of time, $t + 1$ also refers to an arbitrary unit of time (e.g., one day, two weeks, or a month)

Most conditional models for VaR forecast the density directly, although some only attempt to estimate the required quantile of the time $t + 1$ return distribution. Five standard methods are presented in the order of the strength of the assumptions required to justify the method, from strongest to weakest.

¹It is theoretically possible for VaR to be negative. If the VaR of a portfolio is negative, either the portfolio has no risk, the portfolio manager extremely skillful, or most likely the model used to compute the VaR is badly misspecified.

8.3.1 RiskMetrics[®]

The RiskMetrics group has produced a simple, robust method for producing conditional VaR. The basic structure of the RiskMetrics model relies on a restricted GARCH(1,1) where $\alpha + \beta = 1$ and $\omega = 0$. The estimate of the portfolio's variance is

$$\sigma_{t+1}^2 = (1 - \lambda)r_t^2 + \lambda\sigma_t^2, \quad (8.4)$$

where r_t is the (percentage) return on the portfolio in period t . In the RiskMetrics specification σ_{t+1}^2 follows an EWMA which places weight $\lambda^j(1 - \lambda)$ on r_{t-j}^2 .² The RiskMetrics model does not include a conditional mean of the portfolio return, and so is only applicable to assets with returns that are close to zero. The restriction limits the applicability to applications where the risk-measurement horizon is short (e.g., one day to one month). The VaR is constructed from the α -quantile of a normal distribution,

$$\text{VaR}_{t+1} = -\sigma_{t+1}\Phi^{-1}(\alpha) \quad (8.5)$$

where $\Phi^{-1}(\cdot)$ is the inverse normal CDF. The RiskMetrics model has no parameters to estimate; λ has been calibrated to .94 for daily data, 0.97 for weekly data, and .99 for monthly data.³ This model can also be extended to multiple assets using by replacing the squared return with the outer product of a vector of returns, $\mathbf{r}_t \mathbf{r}'_t$, and σ_{t+1}^2 with a matrix, Σ_{t+1} . The limitations of the RiskMetrics model are that the parameters aren't estimated (which is also an advantage), the model does not account for a leverage effect, and the VaR follows a random walk since $\lambda + (1 - \lambda) = 1$.

8.3.2 Parametric ARCH Models

Parametric ARCH-family models provide a complete description of the future return distribution, and so can be applied to estimate the VaR of a portfolio. This model is highly adaptable since the mean, variance and distribution can all be tailored to the portfolio's historical returns. For simplicity, this example uses a constant mean and has a GARCH(1,1) variance process.⁴

$$\begin{aligned} r_{t+1} &= \mu + \varepsilon_{t+1} \\ \sigma_{t+1}^2 &= \omega + \gamma_1 \varepsilon_t^2 + \beta_1 \sigma_t^2 \\ \varepsilon_{t+1} &= \sigma_{t+1} e_{t+1} \\ e_{t+1} &\stackrel{\text{i.i.d.}}{\sim} F(0, 1) \end{aligned}$$

where $F(0, 1)$ is used to indicate that the distribution of innovations need not be normally distributed but must have mean 0 and variance 1. For example, F could be a standardized Student's t with v degrees of freedom or Hansen's skewed t with a degree of freedom parameter v and asymmetry

²An EWMA differs from a standard moving average in two ways. First, an EWMA places relatively more weight on recent observations than on observation in the distant past. Second, EWMA depends on the entire history rather than a fixed-length window.

³The suggested coefficients for λ are based on a large study of the RiskMetrics model across different asset classes.

⁴The use of α_1 in ARCH models has been avoided to avoid confusion with the α in the VaR.

parameter λ . The parameters of the model are estimated using maximum likelihood and the time t conditional VaR is

$$VaR_{t+1} = -\hat{\mu} - \hat{\sigma}_{t+1} F_\alpha^{-1}$$

where F_α^{-1} is the α -quantile of the distribution of e_{t+1} . The flexibility to build a model by specifying the mean, variance and distributions is the strength of this approach. The limitations of this procedure are that implementations require knowledge of a density family which includes F – if the distribution is misspecified then the quantile used is wrong – and that the residuals must come from a location-scale family. The second limitation imposes that all of the dynamics of returns can be summarized by a time-varying mean and variance, and so higher order moments must be time invariant.

8.3.3 Semiparametric ARCH Models/Filtered Historical Simulation

Semiparametric estimation mixes parametric mean and variance models with nonparametric estimators of the distribution.⁵ Again, consider a constant mean GARCH(1,1) model

$$\begin{aligned} r_{t+1} &= \mu + \varepsilon_{t+1} \\ \sigma_{t+1}^2 &= \omega + \gamma_1 \varepsilon_t^2 + \beta_1 \sigma_t^2 \\ \varepsilon_{t+1} &= \sigma_{t+1} e_{t+1} \\ e_{t+1} &\stackrel{\text{i.i.d.}}{\sim} G(0, 1) \end{aligned}$$

where $G(0, 1)$ is an unknown distribution with mean zero and variance 1. Conditional VaR estimates from semiparametric models are also known as filtered Historical Simulation, due to their similarity to Historical Simulation (see). The ARCH model filters the return data by removing the conditional mean and volatility.

When the distribution of the standardized residuals $G(\cdot)$ is unknown, maximum likelihood estimation cannot be used to estimate model parameters. Recall that assuming a normal distribution for the standardized residuals, even if misspecified, produces estimates which are *strongly consistent*, and so ω , γ_1 and β_1 converge to their true values for most any $G(\cdot)$. The model is estimated using QMLE by assuming that the errors are normally distributed and then the Value-at-Risk for the α -quantile can be computed

$$VaR_{t+1}(\alpha) = -\hat{\mu} - \hat{\sigma}_{t+1} \hat{G}_\alpha^{-1} \quad (8.6)$$

where \hat{G}_α^{-1} is the empirical α -quantile of the standardized returns, $\{\hat{e}_{t+1}\}$. To estimate this quantile, define $\hat{e}_{t+1} = \hat{e}_{t+1}/\hat{\sigma}_{t+1}$. and order the errors such that

$$\hat{e}_1 < \hat{e}_2 < \dots < \hat{e}_{n-1} < \hat{e}_n.$$

Here n replaces T to indicate the residuals are no longer time ordered. $\hat{G}_\alpha^{-1} = \hat{e}_{\lfloor \alpha n \rfloor}$ or $\hat{G}_\alpha^{-1} = \hat{e}_{\lceil \alpha n \rceil}$ where $\lfloor x \rfloor$ and $\lceil x \rceil$ denote the floor (largest integer smaller than) and ceiling (smallest integer larger

⁵Semiparametric estimators combine parametric and nonparametric estimators in a single model. In time-series applications, semiparametric estimators have parametric models for the dynamics of the mean and variance but use a nonparametric estimator of the distribution of the residuals.

than) of x .⁶ The estimate of G^{-1} is the α -quantile of the empirical distribution of \hat{e}_{t+1} which is the value in position αn of the ordered standardized residuals.

Semiparametric ARCH models provide one clear advantage over their parametric ARCH cousins; the quantile, and hence the VaR, is consistent under weaker conditions since the density of the standardized residuals does not have to be assumed. The primary disadvantage of the semiparametric approach is that \hat{G}_α^{-1} may be poorly estimated – especially if α is very small (e.g., 1%). Semiparametric ARCH models also share the limitation they are only applicable when returns are generated by a location-scale distribution.

8.3.4 Cornish-Fisher Approximation

The Cornish-Fisher estimator of VaR lies between fully parametric model and the semiparametric model. The setup is identical to that of the semiparametric model,

$$\begin{aligned} r_{t+1} &= \mu + \varepsilon_{t+1} \\ \sigma_{t+1}^2 &= \omega + \gamma \varepsilon_t^2 + \beta \sigma_t^2 \\ \varepsilon_{t+1} &= \sigma_{t+1} e_{t+1} \\ e_{t+1} &\stackrel{\text{i.i.d.}}{\sim} G(0, 1) \end{aligned}$$

where $G(\cdot)$ is an unknown distribution. The model parameters are estimated by QML assuming that the conditional distribution of residuals is normal to produce standardized residuals, $\hat{e}_{t+1} = \hat{\varepsilon}_{t+1}/\hat{\sigma}_{t+1}$. The Cornish-Fisher approximation is a Taylor-series-like expansion of the α -quantile around the α -quantile of a normal and is given by

$$\text{VaR}_{t+1} = -\mu - \sigma_{t+1} G_{CF}^{-1}(\alpha) \quad (8.7)$$

$$\begin{aligned} G_{CF}^{-1}(\alpha) &\equiv \Phi^{-1}(\alpha) + \frac{\varsigma}{6} \left([\Phi^{-1}(\alpha)]^2 - 1 \right) + \\ &\quad \frac{\kappa - 3}{24} \left([\Phi^{-1}(\alpha)]^3 - 3\Phi^{-1}(\alpha) \right) - \frac{\varsigma^2}{36} \left(2[\Phi^{-1}(\alpha)]^3 - 5\Phi^{-1}(\alpha) \right) \end{aligned} \quad (8.8)$$

where ς and κ are the skewness and kurtosis of \hat{e}_{t+1} , respectively. From the expression for $G_{CF}^{-1}(\alpha)$, negative skewness and excess kurtosis ($\kappa > 3$, the kurtosis of a normal) decrease the estimated quantile and increases the VaR. The Cornish-Fisher approximation shares the strength of the semiparametric distribution in that it can be accurate without a parametric assumption. However, unlike the semiparametric estimator, Cornish-Fisher estimators are not necessarily consistent which may be a drawback. Additionally, estimates of higher-order moments of standardized residuals may be problematic or, in very heavy-tailed distributions, the third and fourth moments may not even exist.

⁶When estimating a quantile from discrete data and not smoothing, the is quantile “set valued” and defined as any point between $\hat{e}_{\lfloor \alpha n \rfloor}$ and $\hat{e}_{\lceil \alpha n \rceil}$, inclusive.

8.3.5 Conditional Autoregressive Value-at-Risk (CaViaR)

Engle and Manganelli (2004) developed a family of ARCH-like models to estimate the conditional Value-at-Risk using quantile regression. CaViaR models have a similar structure to GARCH models. The α -quantile of the return distribution, $F_{\alpha,t+1}^{-1}$, is modeled as a weighted average of a constant, the previous value of the quantile, and a shock (or surprise). The shock can take many forms although a “*HIT*”, defined as an exceedance of the previous Value-at-Risk, is the most natural.

$$HIT_{t+1} = I_{[r_{t+1} < F_{\alpha,t+1}^{-1}]} - \alpha \quad (8.9)$$

where r_{t+1} the (percentage) return and $F_{\alpha,t+1}^{-1}$ is the time t α -quantile of this distribution. When $F_{\alpha,t+1}^{-1}$ is the conditional quantile of the return distribution, then a *HIT* is mean zero $E_t \left[I_{[r_{t+1} < F_{\alpha,t+1}^{-1}]} \right] = \Pr(r_{t+1} < F_{\alpha,t+1}^{-1}) = \alpha$.

Defining q_{t+1} as the time $t + 1$ α -quantile of returns, the evolution in a standard CaViaR model is defined by

$$q_{t+1} = \omega + \gamma HIT_t + \beta q_t. \quad (8.10)$$

Other forms that have been explored include the symmetric absolute value,

$$q_{t+1} = \omega + \gamma |r_t| + \beta q_t. \quad (8.11)$$

the asymmetric absolute value,

$$q_{t+1} = \omega + \gamma_1 |r_t| + \gamma_2 |r_t| I_{[r_t < 0]} + \beta q_t \quad (8.12)$$

the indirect GARCH,

$$q_{t+1} = (\omega + \gamma r_t^2 + \beta q_t^2)^{\frac{1}{2}}. \quad (8.13)$$

The parameters of CaViaR models are estimated by minimizing the “tick” loss function

$$\begin{aligned} \arg \min_{\theta} T^{-1} \sum_{t=1}^T & \underbrace{\alpha(r_t - q_t)(1 - I_{[r_t < q_t]})}_{\text{Positive errors}} + \underbrace{(1 - \alpha)(q_t - r_t)I_{[r_t < q_t]}}_{\text{Negative Errors}} = \\ & \arg \min_{\theta} T^{-1} \sum_{t=1}^T \alpha(r_t - q_t) + (q_t - r_t)I_{[r_t < q_t]} \end{aligned} \quad (8.14)$$

where $I_{[r_t < q_t]}$ is an indicator variable which is 1 if $r_t < q_t$ and 0 otherwise. The loss function is linear in the error $r_t - q_t$ and has a slope of α for positive errors and $1 - \alpha$ for negative errors. Estimation of the parameters is complicated since the objective function may be non-differentiable and has many flat spots. Derivative-free methods, such as the Nelder-Mead simplex method or genetic algorithms, can be used to overcome this difficulty. The VaR in a CaViaR framework is then

$$\text{VaR}_{t+1} = -q_{t+1} = -\hat{F}_{t+1}^{-1} \quad (8.15)$$

CaViaR model does not specify a distribution of returns or any moments, and so its use is justified under much weaker assumptions than other VaR estimators. Additionally, its parametric form provides reasonable convergence of the unknown parameters. The main drawbacks of the CaViaR modeling strategy are that it may produce out-of-order quantiles (i.e., 5% VaR is less than 10% VaR) and that estimation of the model parameters is challenging.

8.3.6 Weighted Historical Simulation

Weighted historical simulation constructs an empirical distribution where recent returns are given more weight than returns further in the past. The estimator is nonparametric since that no specific assumptions about either distribution or the dynamics of returns are made.

Weights are assigned using an exponentially declining function. If returns are available from $i = 1, \dots, t$, then the weight given to data point i is

$$w_i = \lambda^{t-i} (1 - \lambda^t), \quad i = 1, 2, \dots, t.$$

Typical values for λ range from .99 to .995. When $\lambda = .99$, 99% of the weight occurs in the most recent 450 data points – .995 changes this to the most recent 900 data points. Smaller values of lambda produce a VaR that is more “local” while larger values produce VaR estimates based most of the historical sample.

The weighted empirical CDF is then

$$\hat{G}_t(r) = \sum_{i=1}^t w_i I_{[r_i < r]}.$$

The conditional VaR is then computed as the solution to

$$\text{VaR}_{t+1} = \min_r \hat{G}(r) \geq \alpha$$

which chooses the smallest value of r where there is at least α probability below in the weighted cumulative distribution.

8.3.7 Example: Conditional Value-at-Risk for the S&P 500

The concepts of VaR is illustrated using S&P 500 returns from January 1, 1999, until December 31, 2018. A variety of models have been estimated that all produce similar VaR estimates. Alternative distributional assumptions generally produce similar volatility parameter estimates in ARCH models, and so VaR estimates only differ due to differences in the quantiles. Table 8.1 reports parameter estimates from these models. The volatility parameters of the TARCH models were virtually identical across all three distributional assumptions. The degree of freedom parameter $\hat{\nu} \approx 8$ in both the standardized Student's t and the skewed t indicating that the standardizes residuals are leptokurtotic, and $\hat{\lambda} \approx -.1$, from the skewed t , indicating some negative skewness. The CaViaR estimates indicate little change in the conditional quantile for positive shock, a substantial increase in the VaR when the return is negative, and that the conditional quantile is highly persistent. The table also contains estimated quantiles using the parametric, semiparametric and Cornish-Fisher expansion of the normal. Since

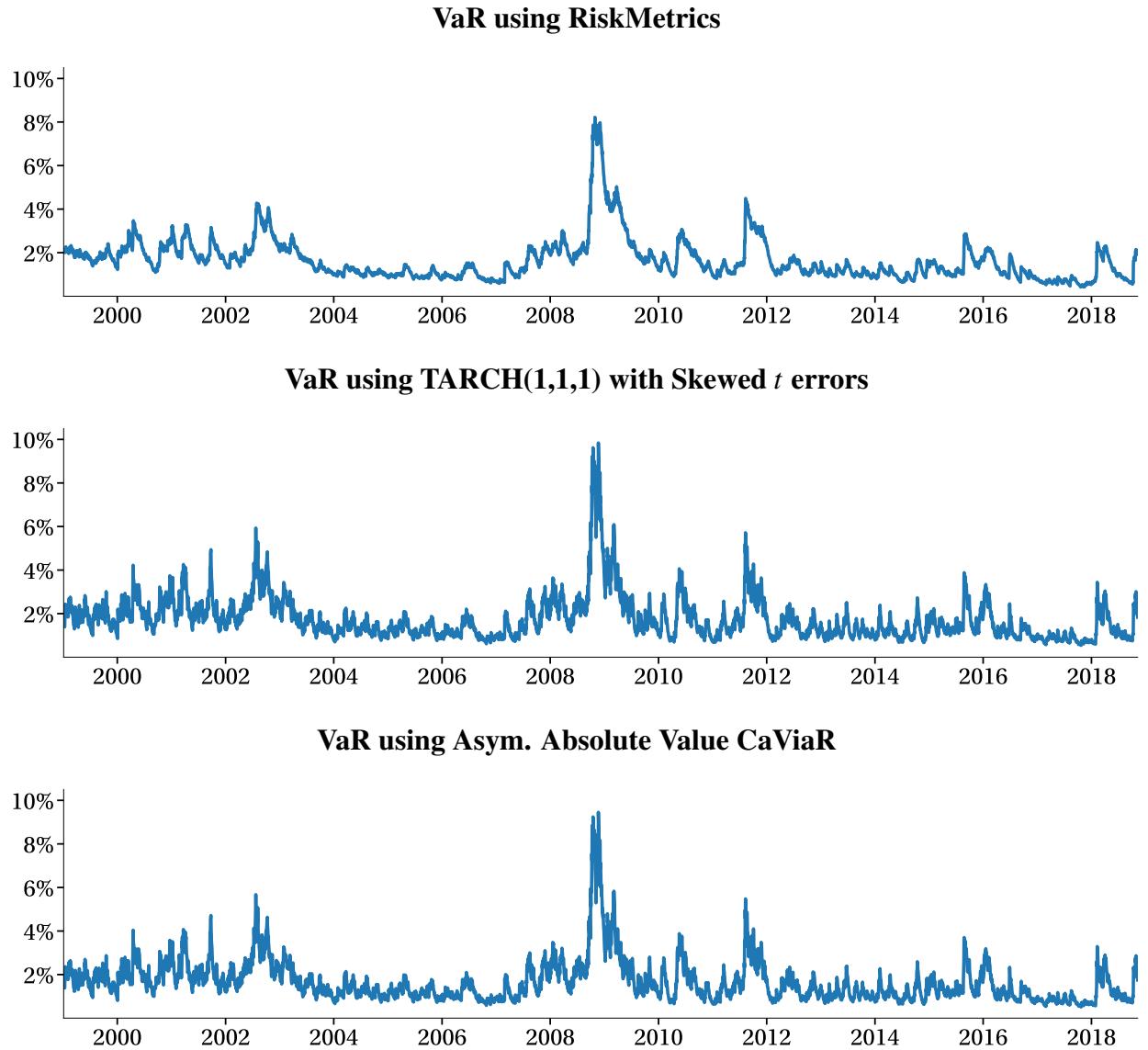


Figure 8.2: The figure contains the estimated 5% VaR for the S&P 500 using data from 1999 until the end of 2018. While these three models have different specifications for the evolution of the conditional VaR, the estimated VaRs are remarkably similar.

the fit conditional variances were nearly identical, the only meaningful difference in the VaRs comes from the differences in these quantiles. These are all qualitatively similar except at 1%.

Figure 8.2 plots the fitted VaRs from the RiskMetrics model, a TARCH with skewed t errors and an asymmetric absolute value CaViaR. All three plots appear very similar, and the TARCH and CaViaR model fits are virtually identical. This similarity is due to the common structure of the dynamics and the values of the estimated parameters. Figure 8.3 plots the conditional VaRs for the weighted Historical Simulation estimator for three values of the smoothing parameter λ . The three values of λ , 0.95, 0.99, and 0.995, places 90% of the weight on the most recent 45, 230 and 2280 observations, respectively. The different values of the smoothing parameter produce meaningfully different condi-

VaR using Weighted Historical Simulation

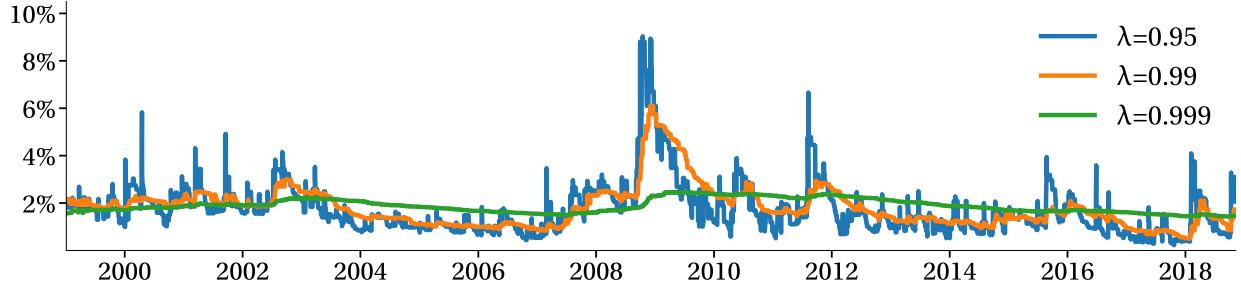


Figure 8.3: The estimated 5% VaR for the S&P 500 using weighted Historical Simulation for $\lambda \in \{0.95, 0.99, 0.999\}$. The three values of λ place 90% of the weight on the most recent 45, 230, and 2280 observations, respectively. Larger values of the decay parameter λ produce smoother conditional VaR estimates.

Model Parameters

	TARCH(1,1,1)					
	ω	γ_1	γ_2	β	ν	λ
Normal	0.026	0.000	0.172	0.909		
Student's t	0.020	0.000	0.173	0.913	7.926	
Skew t	0.022	0.000	0.179	0.910	8.520	-0.123

	CaViaR			
	ω	γ_1	γ_2	β
Asym CaViaR	0.035	0.002	0.290	0.910

Estimated Quantiles from Parametric and Semi-parametric TARCH models

	Semiparam.	Normal	Stud. t	Skew t	CF
1%	-2.656	-2.326	-2.510	-2.674	-2.918
5%	-1.705	-1.645	-1.610	-1.688	-1.739
10%	-1.265	-1.282	-1.209	-1.247	-1.237

Table 8.1: Estimated model parameters and quantiles. The choice of distribution for the standardized shocks makes little difference in the parameters of the TARCH process, and so the fit conditional variances are virtually identical. The only difference in the VaRs from these three specifications comes from the estimates of the quantiles of the standardized returns (bottom panel).

tional VaR estimates. The smallest value appears to produce the conditional VaR estimates are most similar to those depicted in Figure 8.2.

8.4 Unconditional Value at Risk

While the conditional VaR is often the object of interest, there may be situations which call for the unconditional VaR (also known as marginal VaR). Unconditional VaR expands the set of choices from the conditional to include models that do not make use of conditioning information to estimate the VaR directly from the historical return data.

8.4.1 Parametric Estimation

The simplest form of unconditional VaR specifies a complete parametric model for the unconditional distribution of returns. The VaR is then computed from the α -quantile of this distribution. For example, if $r_t \sim N(\mu, \sigma^2)$, then the α -VaR is

$$\text{VaR} = -\mu - \sigma\Phi^{-1}(\alpha). \quad (8.16)$$

The parameters of the distribution are estimated using Maximum likelihood with the usual estimators,

$$\hat{\mu} = T^{-1} \sum_{t=1}^T r_t \quad \hat{\sigma}^2 = T^{-1} \sum_{t=1}^T (r_t - \hat{\mu})^2$$

In a general parametric VaR model, some distribution for returns which depends on a set of unknown parameters θ is assumed, $r_t \sim F(\theta)$ and parameters are estimated by maximum likelihood. The VaR is then $-F_\alpha^{-1}$, where F_α^{-1} is the α -quantile of the estimated distribution. The advantages and disadvantages to parametric unconditional VaR are identical to parametric conditional VaR. The models are parsimonious, and the parameters estimates are precise yet finding a specification general enough to capture the true distribution is difficult.

8.4.2 Nonparametric Estimation/Historical Simulation

At the other end of the spectrum is a simple nonparametric estimate of the unconditional VaR known as Historical Simulation. As was the case in the semiparametric conditional VaR, the first step is to sort the returns so that

$$r_1 < r_2 < \dots < r_{n-1} < r_n$$

where $n = T$ is used to denote an ordering not based on time. The VaR is estimated using $r_{\lfloor \alpha n \rfloor}$ or $r_{\lceil \alpha n \rceil}$ (or any value between the two). The estimate of the VaR is the α -quantile of the empirical distribution of $\{r_t\}$,

$$\text{VaR} = -\hat{G}_\alpha^{-1} \quad (8.17)$$

where \hat{G}_α^{-1} is the estimated quantile. The empirical CDF is defined

$$G(r) = T^{-1} \sum_{t=1}^T I_{[r_t < r]}$$

where $I_{[r_t < r]}$ is an indicator function that takes the value 1 if r_t is less than r , and so this function counts the percentage of returns which are smaller than r .

Historical simulation estimates are rough, and a single new data point may produce very different VaR estimates when the sample size is small. Smoothing the estimated quantile using a kernel density generally improves the precision of the estimate when compared to one calculated directly on the sorted returns. Smoothing the distribution is most beneficial when the sample size is small. See section 8.7.2 for more details.

The advantage of nonparametric estimates of VaR is that they are generally consistent under minimal assumptions about the distribution of returns and that they are trivial to compute. The disadvantage is that the VaR estimates can be poorly estimated – or equivalently that very large samples are needed for estimated quantiles to be accurate – particularly for 1% VaRs (or smaller).

8.4.3 Parametric Monte Carlo

Parametric Monte Carlo is meaningfully different from either parametric or nonparametric estimation of the unconditional distribution. Rather than fit a model to the returns directly, parametric Monte Carlo fits a parsimonious *conditional* model. This model is then used to simulate the *unconditional* distribution. For example, suppose that returns followed an AR(1) with GARCH(1,1) errors and normal innovations,

$$\begin{aligned} r_{t+1} &= \phi_0 + \phi_1 r_t + \varepsilon_{t+1} \\ \sigma_{t+1}^2 &= \omega + \gamma \varepsilon_t^2 + \beta \sigma_t^2 \\ \varepsilon_{t+1} &= \sigma_{t+1} e_{t+1} \\ e_{t+1} &\stackrel{\text{i.i.d.}}{\sim} N(0, 1). \end{aligned}$$

Parametric Monte Carlo is implemented by first estimating the parameters of the model, $\hat{\theta} = [\hat{\phi}_0, \hat{\phi}_1, \hat{\omega}, \hat{\gamma}, \hat{\beta}]'$, and then simulating a long sample $\{\tilde{r}_t\}$ from the process (generally much longer than the actual number of data points available). The VaR from this model is the α -quantile of the *simulated* data.

$$\text{VaR} = -\hat{\tilde{G}}_\alpha^{-1} \quad (8.18)$$

where $\hat{\tilde{G}}_\alpha^{-1}$ is the empirical α -quantile of the simulated data, $\{\tilde{r}_t\}$. Generally, the amount of simulated data should be sufficiently large that the empirical quantile is an accurate estimate of the quantile of the unconditional distribution. There are two advantages to parametric Monte Carlo over other unconditional VaR estimators. First, this procedure exploits *conditioning* information that is ignored by the other estimators. Second, parsimonious conditional models, e.g., ARCH models with leverage, can generate rich families of unconditional distributions that are difficult to parameterize directly. The drawback of this procedure is that an incorrect conditional specification leads to an inconsistent estimate of the unconditional VaR.

8.4.4 Example: Unconditional Value-at-Risk for the S&P 500

Using the S&P 500 data, 3 unconditional parametric models, a normal, a Student's t and a skewed t were estimated. Estimates of the unconditional VaR using the Cornish-Fisher estimator and a Historical Simulation (nonparametric) estimator were also included. Estimates are reported in Table 8.2. The

	Unconditional Value-at-Risk				
	HS	Normal	Stud. t	Skew t	CF
1% VaR	3.313	2.767	3.480	3.684	5.165
5% VaR	1.854	1.949	1.702	1.788	1.754
10% VaR	1.296	1.514	1.150	1.201	0.781

Table 8.2: Unconditional VaR of S&P 500 returns estimated assuming returns are Normal, Student's t or skewed t , using a Cornish-Fisher transformation or using a nonparametric quantile estimator. While the 5% and 10% VaR are similar, the estimates of the 1% VaR differ.

unconditional VaR estimates are similar except for the estimate computed using the Cornish-Fisher expansion. The kurtosis of the data was very high (23) which resulted in a very large 1% quantile. The others are broadly similar with the most substantial differences occurring at the 1% VaR. Figure 8.4 shows the estimated unconditional distribution from the normal and skewed t distributions and a nonparametric kernel density estimator. The key quantiles are similar despite meaningful differences in their shapes.

8.5 Evaluating VaR models

The process of evaluating the performance of VaR models is virtually identical to that of evaluating the specification of models of the conditional mean or variance. The key insight into VaR model evaluation comes from the tick loss function,

$$\sum_{t=1}^T \alpha(r_t - F_{\alpha,t}^{-1})(1 - I_{[r_t < F_{\alpha,t}^{-1}]}) + (1 - \alpha)(F_{\alpha,t}^{-1} - r_t)I_{[r_t < F_{\alpha,t}^{-1}]} \quad (8.19)$$

where r_t is the return in period t and F_t^{-1} is α -quantile of the return distribution in period t . The *generalized error* can be directly computed from this loss function by differentiating with respect to VaR, and is

$$ge_t = I_{[r_t < F_{\alpha,t}^{-1}]} - \alpha \quad (8.20)$$

which is the time- t "HIT" (HIT_t).⁷ When there is a VaR exceedance, $HIT_t = 1 - \alpha$ and when there is no exceedance, $HIT_t = -\alpha$. If the model is correct, then α of the HIT 's should be $(1 - \alpha)$ and $(1 - \alpha)$ should be $-\alpha$, so that

⁷The generalized error extends the concept of an error in a linear regression or linear time-series model to nonlinear estimators. Suppose a loss function is specified as $L(y_{t+1}, \hat{y}_{t+1|t})$, then the generalized error is the derivative of the loss function with respect to the second argument, that is

$$ge_t = \frac{\partial L(y_{t+1}, \hat{y}_{t+1|t})}{\partial \hat{y}_{t+1|t}} \quad (8.21)$$

where it is assumed that the loss function is differentiable at this point.

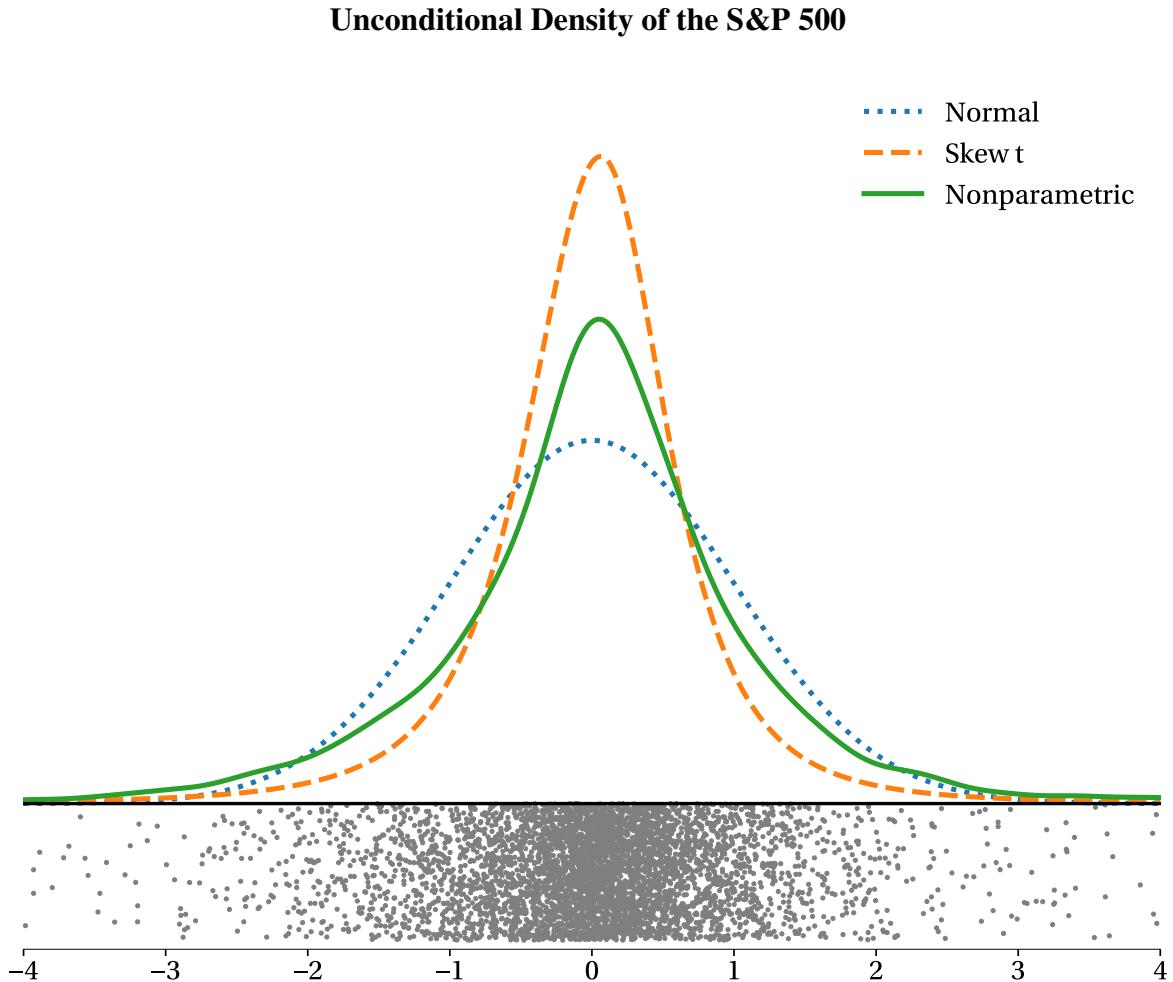


Figure 8.4: Plot of the S&P 500 returns as well as a parametric density using Hansen's skewed t and a nonparametric density estimator constructed using a kernel.

$$\alpha(1 - \alpha) - \alpha(1 - \alpha) = 0,$$

and the mean of HIT_t should be 0. Moreover, when the VaR is conditional on time t information, $E_t[HIT_{t+1}] = 0$ which follows from the properties of optimal forecasts (see chapter 4).

A test that the conditional expectation is zero can be implemented using a generalized Mincer-Zarnowitz (GMZ) regression of $HIT_{t+1|t}$ on any time t available variable. For example, the estimated quantile $F_{t+1|t}^{-1}$ for $t + 1$ could be included (since it is in the time- t information set) as well as lagged HIT s to construct the regression model,

$$HIT_{t+1|t} = \gamma_0 + \gamma_1 F_{t+1|t}^{-1} + \gamma_2 HIT_t + \gamma_3 HIT_{t-1} + \dots + \gamma_K HIT_{t-K+2} + \eta_t$$

If the model is correctly specified, all of the coefficients should be zero and the null $H_0 : \gamma = 0$ can be tested against an alternative that $H_1 : \gamma_j \neq 0$ for some j . If the null is rejected, then either the average number of violations is wrong, so that $\gamma_0 \neq 0$, or the VaR violations are predictable ($\gamma_j \neq 0$ for $j \geq 1$).

8.5.1 Likelihood Evaluation

VaR forecast evaluation can be improved by noting that VaR violations, $I_{[r_t < F_{\alpha,t}^{-1}]}$, are Bernoulli random variables which takes the value 1 with probability α and takes the value 0 with probability $1 - \alpha$. A more powerful test can be constructed using a likelihood ratio test using the Bernoulli random variables $\widetilde{HIT}_t = I_{[r_t < F_{\alpha,t}^{-1}]}$. Under the null that the model is correctly specified, the likelihood function of a series of \widetilde{HIT} s is

$$f(\widetilde{HIT}; p) = \prod_{t=1}^T p^{\widetilde{HIT}_t} (1-p)^{1-\widetilde{HIT}_t}$$

and the log-likelihood is

$$l(p; \widetilde{HIT}) = \sum_{t=1}^T \widetilde{HIT}_t \ln(p) + (1 - \widetilde{HIT}_t) \ln(1-p).$$

If the model is correctly specified, $p = \alpha$ and a likelihood ratio test can be performed as

$$LR = 2(l(\hat{p}; \widetilde{HIT}) - l(p = \alpha; \widetilde{HIT})) \quad (8.22)$$

where $\hat{p} = T^{-1} \sum_{t=1}^T \widetilde{HIT}_t$ is the maximum likelihood estimator of p under the alternative. The test has a single restriction and so has an asymptotic χ_1^2 distribution.

The likelihood-based test for unconditionally correct VaR can be extended to a test of conditionally correct VaR by examining the dependence of HIT s. This testing strategy uses the properties of a Markov chain of Bernoulli random variables. A Markov chain is a modeling device which an ARMA model that models random variables which take on a finite number of values – such as a HIT . A simple 1st order binary valued Markov chain produces Bernoulli random variables which are not necessarily independent. It is characterized by a transition matrix which contains the probability that the state stays the same. In a 1st order binary valued Markov chain, the transition matrix is given by

$$\begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix} = \begin{bmatrix} p_{00} & 1 - p_{00} \\ 1 - p_{11} & p_{11} \end{bmatrix},$$

where p_{ij} is the probability that the next observation takes value j given that this observation has value i . For example, p_{10} indicates that the probability that the next observation is a not a HIT given the current observation is a HIT . In a correctly specified model, the probability of a HIT in the current period should not depend on whether the previous period was a HIT or not. In other words, the sequence $\{\widetilde{HIT}_t\}$ is i.i.d. so that $p_{00} = 1 - \alpha$ and $p_{11} = \alpha$ in a correctly specified model.

Define the following quantities,

$$\begin{aligned} n_{00} &= \sum_{t=1}^{T-1} (1 - \widetilde{HIT}_t)(1 - \widetilde{HIT}_{t+1}) \\ n_{10} &= \sum_{t=1}^{T-1} \widetilde{HIT}_t(1 - \widetilde{HIT}_{t+1}) \end{aligned}$$

$$\begin{aligned} n_{01} &= \sum_{t=1}^{T-1} (1 - \widetilde{HIT}_t) \widetilde{HIT}_{t+1} \\ n_{11} &= \sum_{t=1}^{T-1} \widetilde{HIT}_t \widetilde{HIT}_{t+1} \end{aligned}$$

where n_{ij} counts the number of times $\widetilde{HIT}_{t+1} = i$ after $\widetilde{HIT}_t = j$.

The log-likelihood for the sequence two VaR exceedances is

$$l(p; \widetilde{HIT}) = n_{00} \ln(p_{00}) + n_{01} \ln(1 - p_{00}) + n_{11} \ln(p_{11}) + n_{10} \ln(1 - p_{11})$$

where p_{11} is the probability of two consecutive HIT s and p_{00} is the probability of two sequential periods without a HIT . The null is $H_0 : p_{11} = 1 - p_{00} = \alpha$. The maximum likelihood estimates of p_{00} and p_{11} are

$$\begin{aligned} \hat{p}_{00} &= \frac{n_{00}}{n_{00} + n_{01}} \\ \hat{p}_{11} &= \frac{n_{11}}{n_{11} + n_{10}} \end{aligned}$$

and the null hypothesis can be tested using the likelihood ratio

$$LR = 2(l(\hat{p}_{00}, \hat{p}_{11}; \widetilde{HIT}) - l(p_{00} = 1 - \alpha, p_{11} = \alpha; \widetilde{HIT})). \quad (8.23)$$

This test has an asymptotic χ^2_2 distribution since there are two restrictions under the null.

This framework can be extended to include conditioning information by specifying a *probit* or *logit* for \widetilde{HIT}_t using any time- t available information. Both of these models are known as limited dependent variable models since the left-hand-side variables are always 0 or 1. For example, a specification test could be constructed using K lags of HIT , a constant and the forecast quantile as

$$\widetilde{HIT}_{t+1|t} = \gamma_0 + \gamma_1 F_{t+1|t} + \gamma_2 \widetilde{HIT}_t + \gamma_3 \widetilde{HIT}_{t-1} + \dots + \gamma_K \widetilde{HIT}_{t-K+1}.$$

Parameters are computed by maximizing the Bernoulli log-likelihood, which requires the estimated probabilities to satisfy

$$0 \leq \gamma_0 + \gamma_1 F_{t+1|t} + \gamma_2 \widetilde{HIT}_t + \gamma_3 \widetilde{HIT}_{t-1} + \dots + \gamma_K \widetilde{HIT}_{t-K+1} \leq 1.$$

This restriction is imposed using one of two transformations, the normal CDF ($\Phi(z)$) which produces the probit model or the logistic function ($e^z/(1 + e^z)$) which produces the logit model. Generally the choice between these two makes little difference. If $\mathbf{x}_t = [1 \ F_{t+1|t} \ \widetilde{HIT}_t \ \widetilde{HIT}_{t-1} \ \dots \ \widetilde{HIT}_{t-K+1}]$, the model for \widetilde{HIT} is

$$\widetilde{HIT}_{t+1|t} = \Phi(\mathbf{x}_t \boldsymbol{\gamma})$$

where the normal CDF is used to map from $(-\infty, \infty)$ to $(0,1)$, and so the model is a conditional probability model. The log-likelihood is

$$l(\gamma; \widetilde{HIT}, \mathbf{x}) = \sum_{t=1}^T \widetilde{HIT}_t \ln(\Phi(\mathbf{x}_t \gamma)) - (1 - \widetilde{HIT}_t) \ln(1 - \Phi(\mathbf{x}_t \gamma)). \quad (8.24)$$

The likelihood ratio for testing the null $H_0 : \gamma_0 = \Phi^{-1}(\alpha), \gamma_j = 0$ for all $j = 1, 2, \dots, K$ against an alternative $H_1 = \gamma_0 \neq \Phi^{-1}(\alpha)$ or $\gamma_j \neq 0$ for some $j = 1, 2, \dots, K$ can be computed

$$LR = 2 \left(l(\hat{\gamma}; \widetilde{HIT}) - l(\gamma_0; \widetilde{HIT}) \right) \quad (8.25)$$

where γ_0 is the value under the null ($\gamma = \mathbf{0}$) and $\hat{\gamma}$ is the estimator under the alternative (i.e., the unrestricted estimator from the probit).

8.5.2 Relative Comparisons

Diebold-Mariano tests can be used to rank the relative performance of VaR forecasting models (Diebold and Mariano, 1995). DM tests of VaR models are virtually identical to DM tests on the forecasts from two conditional mean or conditional variance models. The only important difference is the use of the VaR-specific tick loss function. If $L(r_{t+1}, \text{VaR}_{t+1|t})$ is a loss function defined over VaR, then a Diebold-Mariano test statistic can be computed

$$DM = \frac{\bar{d}}{\sqrt{\widehat{\text{V}[d]}}} \quad (8.26)$$

where

$$d_t = L(r_{t+1}, \text{VaR}_{t+1|t}^A) - L(r_{t+1}, \text{VaR}_{t+1|t}^B),$$

VaR^A and VaR^B are the Value-at-Risks from models A and B respectively, $\bar{d} = R^{-1} \sum_{t=M+1}^{M+R} d_t$, M (for modeling) is the number of observations used in the model building and estimation, R (for reserve) is the number of observations held back for model evaluation, and $\sqrt{\widehat{\text{V}[d]}}$ is the long-run variance of d_t which requires the use of a HAC covariance estimator (e.g., Newey-West). Recall that DM is asymptotically normally distributed. The null is equal accuracy, $H_0 : E[d_t] = 0$, and the composite alternative is $H_1^A : E[d_t] < 0$ and $H_1^B : E[d_t] > 0$. Large negative values (less than -2) indicate model A is superior while large positive values indicate the opposite; values close to zero indicate neither forecasting model outperforms the other.

Ideally the loss function, $L(\cdot)$, should reflect the user's preference over VaR forecast errors. In some circumstances there is no obvious choice, and the tick loss function,

$$L(r_{t+1}, \text{VaR}_{t+1|t}) = \alpha(r_{t+1} - \text{VaR}_{t+1|t})(1 - I_{[r_{t+1} < \text{VaR}_{t+1|t}]}) + (1 - \alpha)(\text{VaR}_{t+1|t} - r_{t+1})I_{[r_{t+1} < \text{VaR}_{t+1|t}]} \quad (8.27)$$

is a theoretically sound choice. When the distribution of returns is continuous, the tick-loss is uniquely minimized at the conditional quantile. The tick-loss function has the same interpretation in a VaR model as the mean square error (MSE) does in conditional mean model evaluation or the QLIK loss function in volatility models evaluation.

8.6 Expected Shortfall

Expected shortfall – also known as tail VaR – combines aspects of VaR with additional information about the distribution of returns in the tail.⁸

Definition 8.4 (Expected Shortfall). Expected Shortfall (ES) is defined as the expected value of the portfolio loss *given* a Value-at-Risk exceedance has occurred. The *unconditional* Expected Shortfall is defined

$$\begin{aligned} \text{ES} &= E \left[\frac{W_{t+1} - W_t}{W_t} \middle| \frac{W_{t+1} - W_t}{W_t} < -\text{VaR} \right] \\ &= E[r_{t+1} | r_{t+1} < -\text{VaR}] \end{aligned} \quad (8.28)$$

where W_t , is the value of the assets in the portfolio.⁹

The conditional, and generally more useful, Expected Shortfall is similarly defined.

Definition 8.5 (Conditional Expected Shortfall). Conditional Expected Shortfall is defined

$$\text{ES}_{t+1} = E_t [r_{t+1} | r_{t+1} < -\text{VaR}_{t+1}]. \quad (8.29)$$

where r_{t+1} return on a portfolio at time $t + 1$. Since t is an arbitrary measure of time, $t + 1$ also refers to an arbitrary unit of time (day, two-weeks, 5 years, etc.)

Because the computation of Expected Shortfall requires both a quantile and an expectation, they are generally computed from density models, either parametric or semiparametric, rather than models focused on only the ES.

8.6.1 Evaluating Expected Shortfall models

Expected Shortfall models can be evaluated using standard techniques since Expected Shortfall is a conditional mean,

$$E_t[\text{ES}_{t+1}] = E_t[r_{t+1} | r_{t+1} < -\text{VaR}_{t+1}].$$

A generalized Mincer-Zarnowitz regression can be used to test whether this mean is zero. Let $I_{[r_t < \text{VaR}_t]}$ indicate that the portfolio return was less than the VaR. The GMZ regression for testing Expected Shortfall is

$$(\text{ES}_{t+1|t} - r_{t+1})I_{[r_{t+1} < -\text{VaR}_{t+1|t}]} = \mathbf{x}_t \gamma \quad (8.30)$$

⁸Expected Shortfall is a special case of a broader class of statistics known as *exceedance measures*. Exceedance measures all describe a common statistic *conditional* on one or more variables being in their tail. Expected shortfall it is an *exceedance mean*. Other exceedance measures which have been studied include exceedance variance, $V[X|X < q_\alpha]$, exceedance correlation, $\text{Corr}(X, Y|X < q_{\alpha,X}, Y < q_{\alpha,Y})$, and exceedance β , $\text{Cov}(X, Y|X < q_{\alpha,X}, Y < q_{\alpha,Y}) / (V[X|X < q_{\alpha,X}]V[Y|Y < q_{\alpha,Y}])^{\frac{1}{2}}$ where $q_{\alpha,\cdot}$ is the α -quantile of the distribution of X or Y .

⁹Just like VaR, Expected Shortfall can be equivalently defined in terms of returns or in terms of wealth. For consistency with the VaR discussion, Expected Shortfall is presented here using the return.

where \mathbf{x}_t , as always, is any set of time t measurable instruments. The natural choices for \mathbf{x}_t include a constant and $ES_{t+1|t}$, the forecast Expected Shortfall. Any other time- t measurable regressors that capture important characteristics of the tail, such as recent volatility or the VaR forecast (VaR_{t+1}), may also be useful in evaluating Expected Shortfall models. If the Expected Shortfall model is correct, the null that none of the regressors are useful in predicting the difference, $H_0 : \gamma = \mathbf{0}$, should not be rejected. If the left-hand side term – the Expected Shortfall “surprise” – in eq. (8.30) is predictable, then the model can be improved.

Despite the simplicity of the GMZ regression framework to evaluate Expected Shortfall, their evaluation is difficult due to the scarcity of data available to evaluate the exceedance mean; Expected Shortfall can only be measured when there is a VaR exceedance and so 4 years of data would only produce 50 observations where this was true. The lack of data about the tail makes evaluating Expected Shortfall models difficult and can lead to a failure to reject in many cases even when using misspecified Expected Shortfall models.

8.7 Density Forecasting

Value-at-Risk, a quantile, provides a narrow view into the riskiness of an asset. More importantly, VaR may not adequately describe the types of risk relevant to a forecast consumer. A density forecast, in contrast, summarizes *everything* there is to know about the riskiness of the asset. Density forecasts nest both VaR and Expected Shortfall as special cases.

In light of this relationship, it is not apparent that VaR or Expected Shortfall should be used. Density forecasting suffers from three distinct challenges:

- The density contains all of the information about the random variable being studied, and so a flexible form is generally needed. The cost of this flexibility is increased parameter estimation error which can be magnified when computing the expectation of nonlinear functions of a forecast density of future asset prices (e.g., pricing an option).
- Multi-step density forecasts are rarely analytically tractable since densities do not time aggregate, except in special cases that are too simple for most applications.
- Unless the user has preferences over the entire distribution, density forecasts inefficiently utilize information.

8.7.1 Density Forecasts from ARCH models

Density forecasting from ARCH models is identical to VaR forecasting from ARCH models. For simplicity, a model with a constant mean and GARCH(1,1) variances is used,

$$\begin{aligned} r_{t+1} &= \mu + \varepsilon_{t+1} \\ \sigma_{t+1}^2 &= \omega + \gamma_1 \varepsilon_t^2 + \beta_1 \sigma_t^2 \\ \varepsilon_{t+1} &= \sigma_{t+1} e_{t+1} \\ e_{t+1} &\stackrel{\text{i.i.d.}}{\sim} G(0, 1). \end{aligned}$$

where $G(0, 1)$ is used to indicate that the distribution of innovations need not be normal but must have mean 0 and variance 1. In practice, the mean and variance can be modeled using richer parameterizations that have been tailored so the historical data. Standard choices for $G(\cdot)$ include the standardized Student's t , the generalized error distribution, and Hansen's skewed t . The 1-step ahead density forecast is then

$$\hat{F}_{t+1|t} \stackrel{d}{=} G(\hat{\mu}, \hat{\sigma}_{t+1|t}^2) \quad (8.31)$$

where $F(\cdot)$ is the distribution of returns. This follows directly from the original model where $r_{t+1} = \mu + \sigma_{t+1} e_{t+1}$ and $e_{t+1} \stackrel{\text{i.i.d.}}{\sim} G(0, 1)$.

8.7.2 Semiparametric Density forecasting

Semiparametric density forecasting is also similar to its VaR counterpart. The model begins by assuming that innovations are generated according to some unknown distribution $G(\cdot)$,

$$\begin{aligned} r_{t+1} &= \mu + \varepsilon_{t+1} \\ \sigma_{t+1}^2 &= \omega + \gamma_1 \varepsilon_t^2 + \beta_1 \sigma_t^2 \\ \varepsilon_{t+1} &= \sigma_{t+1} e_{t+1} \\ e_{t+1} &\stackrel{\text{i.i.d.}}{\sim} G(0, 1). \end{aligned}$$

and estimates of $\hat{\sigma}_t^2$ are computed assuming that the innovations are conditionally normal. The justification for this choice follows from the strong consistency of the variance parameter estimates even when the innovations are not normal. Using the estimated variances, standardized innovations are computed as $\hat{e}_t = \hat{\varepsilon}_t / \hat{\sigma}_t$. The final step is to compute the distribution. The simplest method to accomplish this is to compute the empirical CDF as

$$G(e) = T^{-1} \sum_{t=1}^T I_{[\hat{e}_t < e]}. \quad (8.32)$$

The function returns the percentage of the standardized residuals smaller than the value e . This method is trivial but has some limitations. First, the PDF does not exist since $G(\cdot)$ is not differentiable. This property makes some applications difficult, although a histogram provides a simple, but imprecise, method to work around the non-differentiability of the empirical CDF. Second, the CDF is jagged and is generally an inefficient estimator, particularly in the tails.

An alternative, more accurate estimator can be constructed using a kernel to smooth the density. A kernel density is a local average of the number of \hat{e}_t in a small neighborhood of e . The more standardized residuals in this neighborhood, the higher the probability in the region, and the larger the value of the kernel density. The kernel density estimator is defined

$$g(e) = \frac{1}{Th} \sum_{t=1}^T K\left(\frac{\hat{e}_t - e}{h}\right) \quad (8.33)$$

where $K(\cdot)$ can be one of many kernels – the choice of which usually makes little difference. The two most common are the Gaussian,

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2), \quad (8.34)$$

and the Epanechnikov,

$$K(x) = \begin{cases} \frac{3}{4}(1-x^2) & -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}. \quad (8.35)$$

The choice of the bandwidth h determines the width of the window used to smooth the density. It plays a more substantial role than the choice of the kernel in the accuracy of a density estimate. In practice, Silverman's bandwidth,

$$h = 1.06\sigma T^{-\frac{1}{5}}, \quad (8.36)$$

is widely used where σ is the standard deviation of $\hat{\epsilon}_t$ (which is theoretically 1, but may differ if the model is misspecified). However, larger or smaller bandwidths can be used to produce smoother or rougher densities, respectively. The magnitude of the bandwidth represents a bias-variance tradeoff – a small bandwidth has little bias but is very jagged (high variance), while a large bandwidth produces an estimate with substantial bias but very smooth (low variance). If the CDF is needed, $g(e)$ can be integrated using numerical techniques such as a trapezoidal approximation to the Riemann integral.

Finally, the density forecast is constructed by scaling the distribution G by $\sigma_{t+1|t}$ and adding the mean. The top panel of Figure 8.5 contains a plot of the empirical CDF and kernel smoothed CDF of TARCH(1,1,1)-standardized S&P 500 returns in 2018. The empirical CDF is jagged, and there are some large gaps in the observed returns. The bottom panel shows the histogram of the standardized returns where each bin contains 10 returns, and the smoothed kernel density estimate computed using Silverman's bandwidth and a Gaussian kernel.

8.7.3 Multi-step density forecasting and the fan plot

Multi-step ahead density forecasts do not time aggregate. For example, consider a simple GARCH(1,1) model with normal innovations,

$$\begin{aligned} r_{t+1} &= \mu + \varepsilon_{t+1} \\ \sigma_{t+1}^2 &= \omega + \gamma_1 \varepsilon_t^2 + \beta_1 \sigma_t^2 \\ \varepsilon_{t+1} &= \sigma_{t+1} e_{t+1} \\ e_{t+1} &\stackrel{\text{i.i.d.}}{\sim} N(0, 1). \end{aligned}$$

The 1-step ahead density forecast of returns is

$$r_{t+1} | \mathcal{F}_t \sim N(\mu, \sigma_{t+1|t}^2). \quad (8.37)$$

Since innovations are conditionally normal and $E_t [\sigma_{t+2|t}^2]$ is simple to compute, it is tempting construct a 2-step ahead forecast also using a normal,

$$r_{t+2} | \mathcal{F}_t \sim N(\mu, \sigma_{t+2|t}^2). \quad (8.38)$$

This forecast is not correct since the 2-step ahead distribution is a *variance-mixture* of normals and so is itself non-normal. This reason for the difference is that $\sigma_{t+2|t}^2$, unlike $\sigma_{t+1|t}^2$, is a random variable

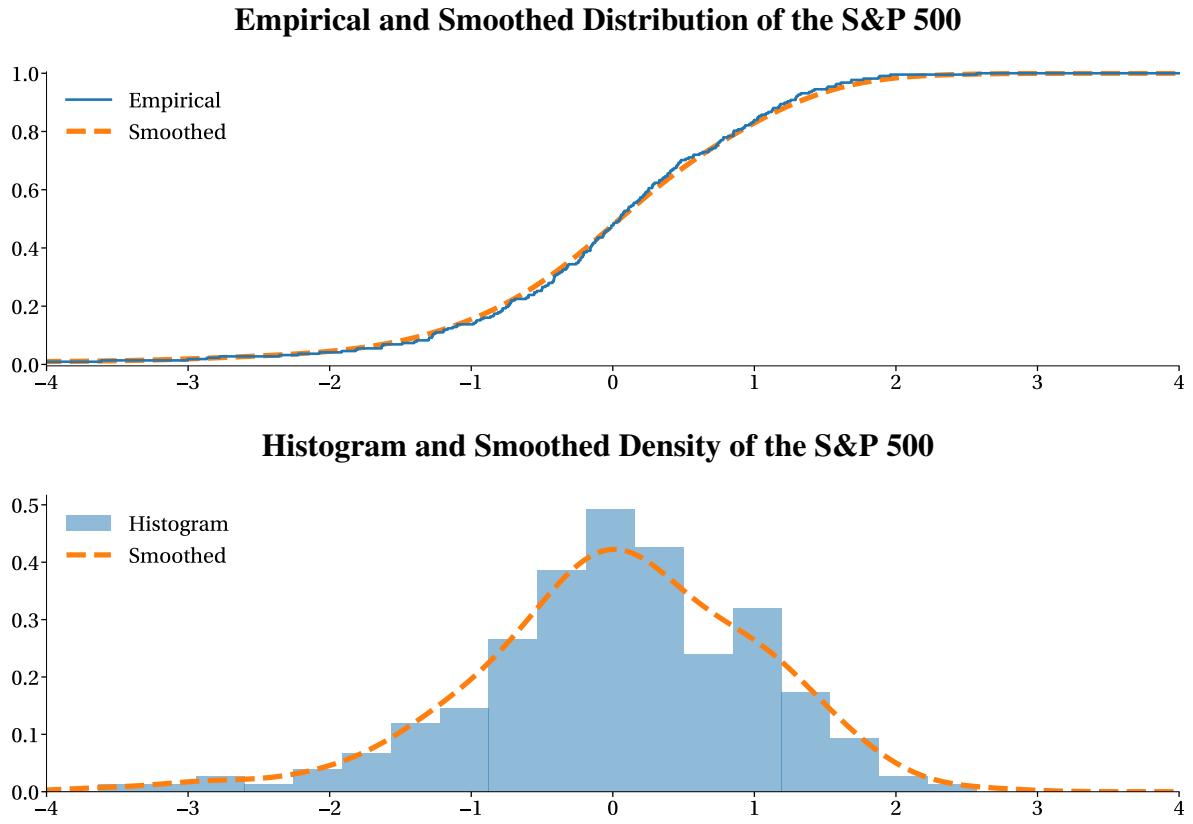


Figure 8.5: The top panel shows the rough empirical and smoothed empirical CDF for standardized returns of the S&P 500 in 2018 (standardized by a TARCH(1,1,1)). The bottom panel shows the histogram of the standardized returns using bins with 10 observations each and the smoothed kernel density.

and the uncertainty in $\sigma_{t+2|t}^2$ must be integrated out to determine the distribution of r_{t+2} . The correct form of the 2-step ahead density forecast is

$$r_{t+2}|\mathcal{F}_t \sim \int_{-\infty}^{\infty} \phi(\mu, \sigma^2(e_{t+1})_{t+2|t+1}) \phi(e_{t+1}) de_{t+1}.$$

where $\phi(\cdot)$ is a normal probability density function and $\sigma^2(e_{t+1})_{t+2|t+1}$ reflects the explicit dependence of $\sigma_{t+2|t+1}^2$ on e_{t+1} . While this expression is fairly complicated, a simpler way to view it is as a mixture of normal random variables where the probability of getting a specific normal depends on $w(e) = \phi(e_{t+1})$,

$$r_{t+2}|\mathcal{F}_t \sim \int_{-\infty}^{\infty} w(e) f(\mu, \sigma(e_{t+1})_{t+2|t+1}) de.$$

Unless $w(e)$ is constant, the resulting distribution is not a normal. The top panel in Figure 8.6 contains the naïve 10-step ahead forecast and the correct 10-step ahead forecast for a simple GARCH(1,1) process,

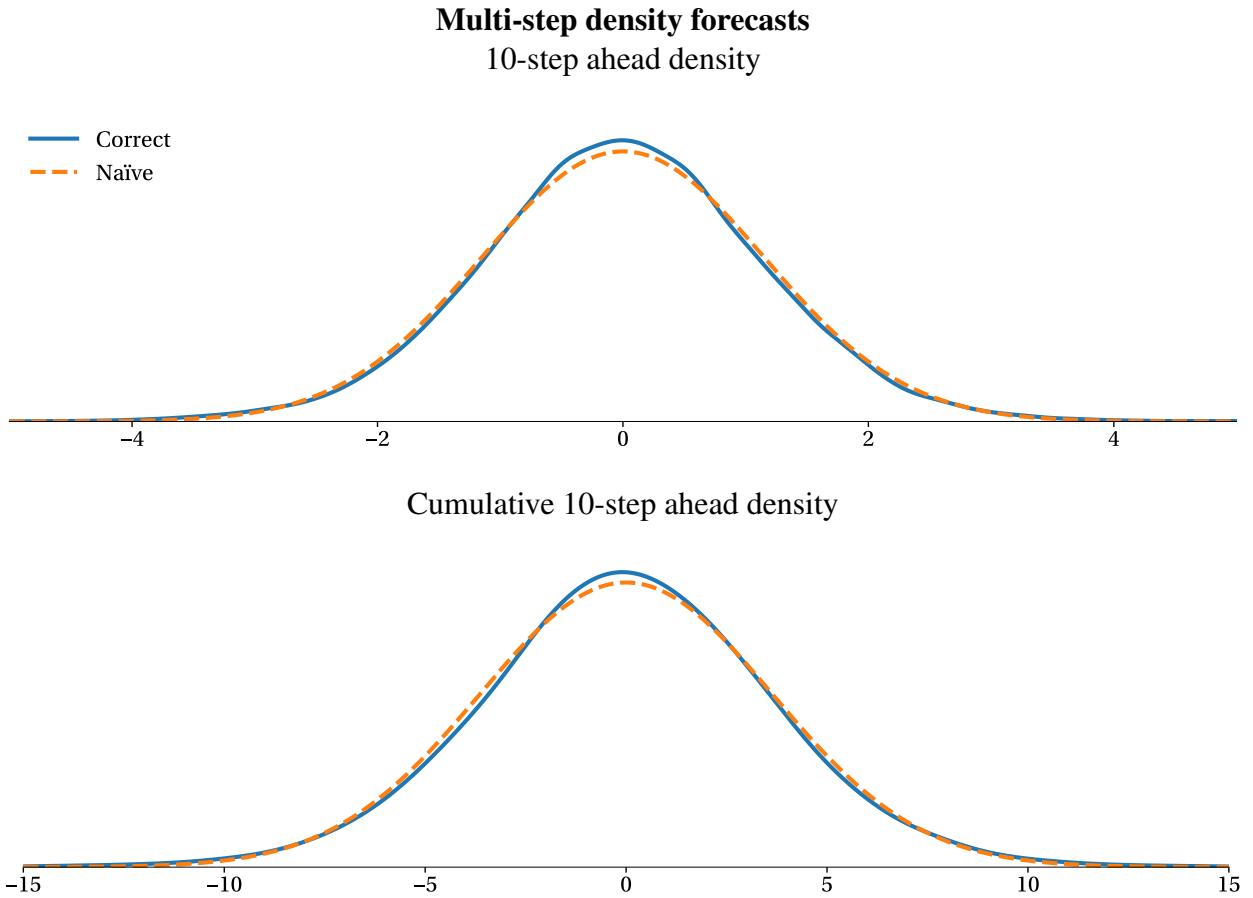


Figure 8.6: Naïve and correct 10-step ahead density forecasts from a simulated GARCH(1,1) model. The correct density forecasts have substantially fatter tails than the naïve forecast as evidenced by the central peak and cross-over of the density in the tails.

$$\begin{aligned}
 r_{t+1} &= \varepsilon_{t+1} \\
 \sigma_{t+1}^2 &= .02 + .2\varepsilon_t^2 + .78\sigma_t^2 \\
 \varepsilon_{t+1} &= \sigma_{t+1} e_{t+1} \\
 e_{t+1} &\stackrel{\text{i.i.d.}}{\sim} N(0, 1)
 \end{aligned}$$

where $\varepsilon_t^2 = \sigma_t^2 = \bar{\sigma} = 1$ and hence $E_t[\sigma_{t+h}] = 1$ for all h . The bottom panel contains the plot of the density of a cumulative 10-day return (the sum of the 10 1-day returns). In this case the naïve model assumes that

$$r_{t+h} | \mathcal{F}_t \sim N(\mu, \sigma_{t+h|t})$$

for $h = 1, 2, \dots, 10$. The correct forecast has heavier tails than the naïve forecast which can be verified by checking that the solid line is above the dashed line in the extremes.

Fan plot of the forecasts of an AR(2)

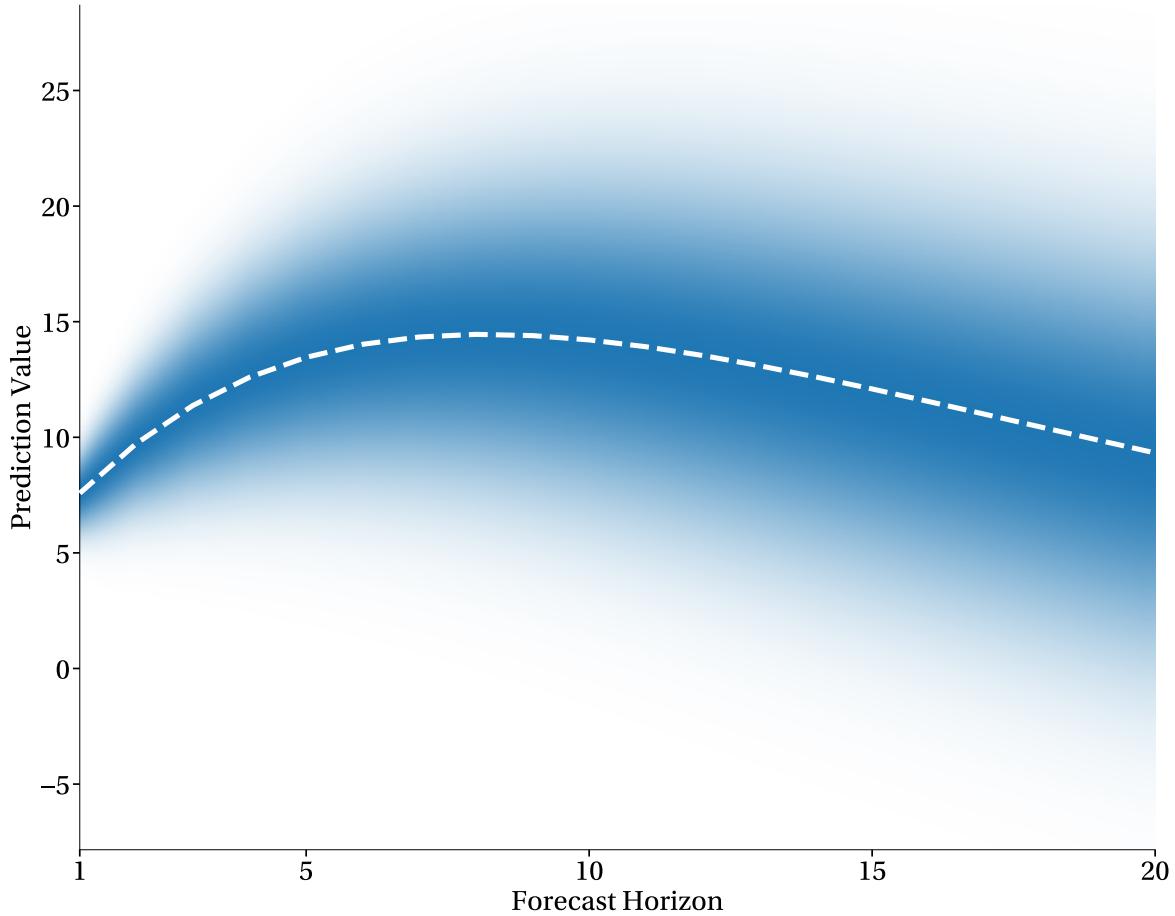


Figure 8.7: Future density of a persistent AR(2) with i.i.d. standard normal increments. Darker regions indicate higher probability while progressively lighter regions indicate less likely events.

Fan Plots

A fan plot is a graphical tool to convey information about future *changes* in uncertainty. The Bank of England has popularized these representations as a method to convey the uncertainty about the future of the economy. Figure 8.7 contains a fan plot of the forecast density for a persistent AR(2) with i.i.d. standard normal increments.¹⁰ Darker regions indicate higher probability while progressively lighter regions indicate less likely events.

8.7.4 Quantile-Quantile (QQ) plots

A Quantile-Quantile, or QQ, plot is a graphical tool that is used to assess the fit of a density or a density forecast. Suppose a set of standardized residuals \hat{e}_t are assumed to have a distribution F . The

¹⁰The density was generated from the AR(2) $Y_t = 1.8Y_{t-1} - 0.81Y_{t-2} + \varepsilon_t$ where the final two in-sample values are $y_T = 4.83$ and $y_{T-1} = 1.37$.

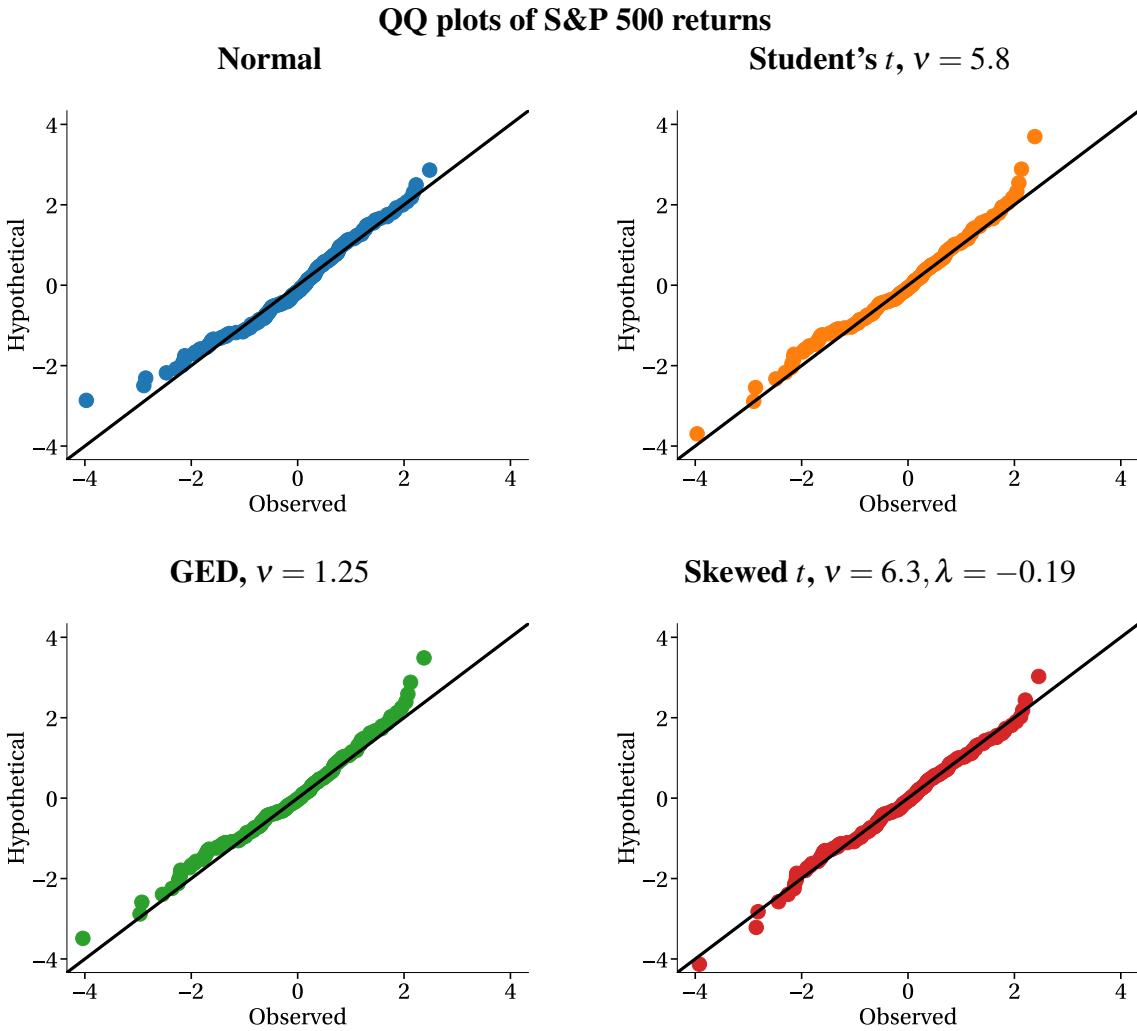


Figure 8.8: QQ plots of the studentized S&P 500 returns against fitted Normal, Student's t , GED and Skewed t distributions. Points along the 45° indicate a good distributional fit.

QQ plot is generated by ordering the standardized residuals,

$$\hat{e}_1 < \hat{e}_2 < \dots < \hat{e}_{n-1} < \hat{e}_n$$

and then plotting the ordered residual \hat{e}_j (x-axis) against its hypothetical value (y-axis) if the correct distribution were F , which is the inverse CDF evaluated at $\frac{j}{T+1}$, $\left(F^{-1}\left(\frac{j}{T+1}\right)\right)$. This graphical assessment of a distribution is formalized in the Kolmogorov-Smirnov test. Figure 8.8 contains 4 QQ plots for monthly S&P 500 returns against a normal, a Student's t , a skewed t , and a GED. The MLE of the density parameters were used to produce the QQ plots. The normal appears to be badly misspecified in the tails – as evidenced through deviations from the 45° line. The other models appear adequate for the monthly returns. The skewed t performs especially well in the lower tail.

8.7.5 Evaluating Density Forecasts

All density evaluation strategies are derived from a basic property of continuous random variables: if $x \sim F$, then $u \equiv F(x) \sim U(0, 1)$. That is, for any continuous random variable X , the cumulant of X has a Uniform distribution over $[0, 1]$. The opposite of this results is also true, if $U \sim \text{Uniform}(0, 1)$, $F^{-1}(U) = X \sim F$.¹¹

Theorem 8.1 (Probability Integral Transform). *Let a random variable X have a continuous, increasing CDF $F_X(x)$ and define $Y = F_X(X)$. Then Y is uniformly distributed and $\Pr(Y \leq y) = y$, $0 < y < 1$.*

Theorem 8.1. For any $y \in (0, 1)$, $Y = F_X(X)$, and so

$$\begin{aligned} F_Y(y) &= \Pr(Y \leq y) = \Pr(F_X(X) \leq y) \\ &= \Pr(F_X^{-1}(F_X(X)) \leq F_X^{-1}(y)) && \text{Since } F_X^{-1} \text{ is increasing} \\ &= \Pr(X \leq F_X^{-1}(y)) && \text{Invertible since strictly increasing} \\ &= F_X(F_X^{-1}(y)) && \text{Definition of } F_X \\ &= y \end{aligned}$$

□

The proof shows that $\Pr(F_X(X) \leq y) = y$ and so $Y = F_X(X)$ must be a uniform random variable (by definition).

The Kolmogorov-Smirnov (KS) test exploits this property of residuals from the correct distribution to test whether a set of observed data are compatible with a specified distribution F . The test statistic is calculated by first computing the *probability integral transformed residuals* $\hat{u}_t = F(\hat{e}_t)$ from the standardized residuals and then sorting them

$$u_1 < u_2 < \dots < u_{n-1} < u_n.$$

The KS test statistic is then computed from

$$\begin{aligned} KS &= \max_{\tau} \left| \sum_{i=1}^{\tau} I_{[u_i < \frac{\tau}{T}]} - \frac{1}{T} \right| \\ &= \max_{\tau} \left| \left(\sum_{i=1}^{\tau} I_{[u_i < \frac{\tau}{T}]} \right) - \frac{\tau}{T} \right| \end{aligned} \tag{8.39}$$

The test statistic finds the maximum deviation between the number of u_j less than $\frac{\tau}{T}$ and the expected number of observations which should be less than $\frac{\tau}{T}$. Since the probability integral transformed residuals should be $\text{Uniform}(0, 1)$ when the model is correctly specified, the number of probability integral transformed residuals expected to be less than $\frac{\tau}{T}$ is $\frac{\tau}{T}$. The distribution of the KS test is nonstandard and simulated critical values are available in most software packages.

¹¹The latter result can be used as the basis of a random number generator. To generate a random number with a CDF of F , first generate a uniform value, u , and then compute the inverse CDF of u to produce a random number from F , $y = F^{-1}(u)$. If the inverse CDF is not available in closed form, monotonicity of the CDF allows for quick, precise numerical inversion.

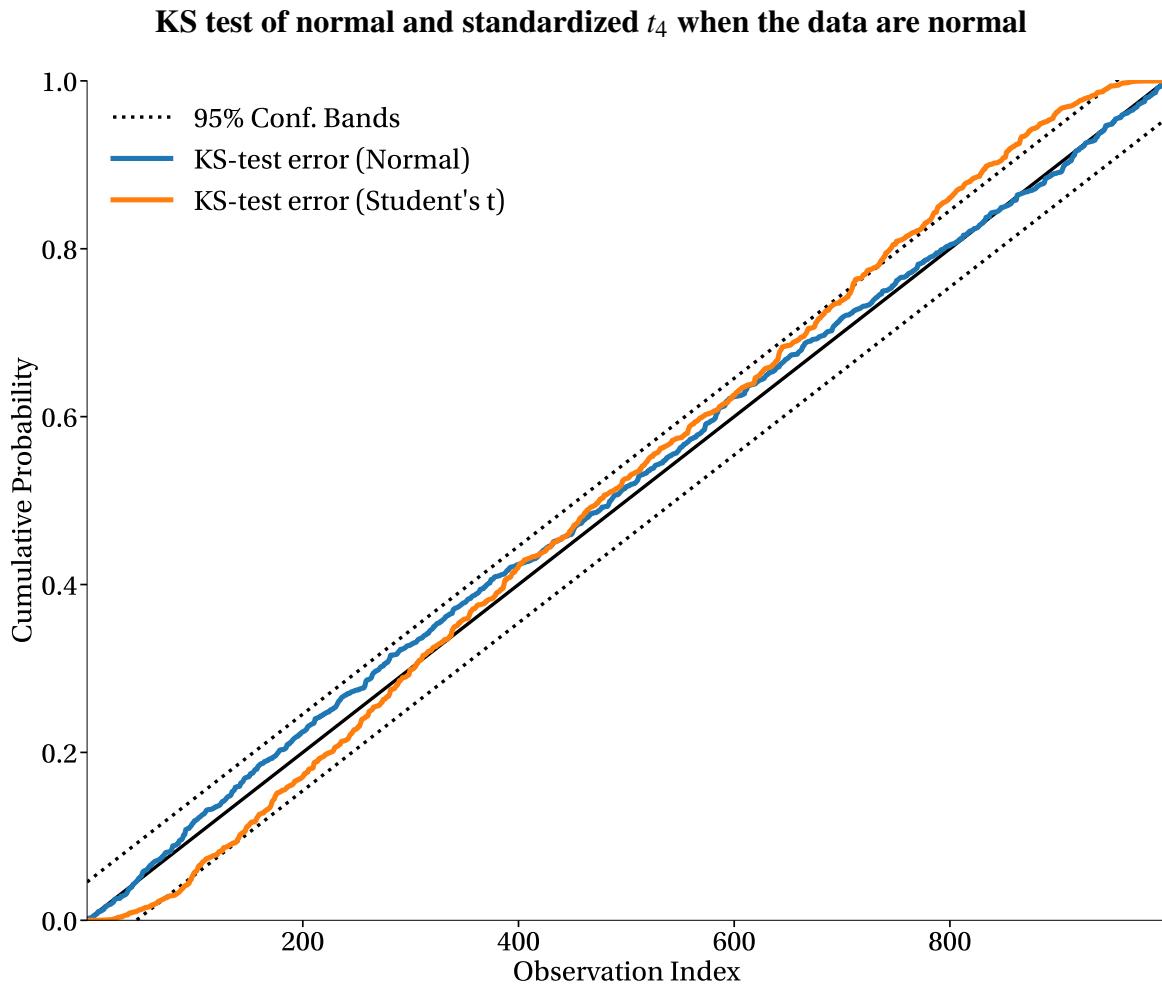


Figure 8.9: A KS test with simulated normal and t_3 data. In both cases, the null is that the data have normal distributions. The data generated from the t_3 crosses the confidence boundary indicating a rejection of this specification. An accurate density forecast should produce a cumulative distribution close to the 45° line.

The KS test has a graphical interpretation as a QQ plot of the probability integral transformed residuals against a uniform. Figure 8.9 contains a representation of the KS test using data from two series: the first is standard normal and the second is a standardized Student's t_3 . 95% confidence bands are denoted with dotted lines. The data from both series were assumed to be from a standard normal. The KS test rejects normality of the t_3 data as evidenced by the cumulants just crossing the confidence band.

Parameter Estimation Error and the KS Test

The critical values supplied by most packages *do not* account for parameter estimation error. KS tests on in-sample data from models with estimated parameters are *less* likely to reject than if the true parameters are known. For example, if a sample of 1000 random variables are i.i.d. standard normal

and the mean and variance are known to be 0 and 1, the 90, 95 and 99% CVs for the KS test are 0.0387, 0.0428, and 0.0512. If the parameters are not known and must be estimated, the 90, 95 and 99% CVs are reduced to 0.0263, 0.0285, 0.0331. Thus, the desired size of 10% (corresponding to a 90% critical value) has an actual size closer to 0.1%. Using the wrong critical value distorts the size of the test and lowers the test's power – the test statistic is unlikely to reject the null hypothesis in many instances where it should.

The solution to this problem is simple. Since the KS-test requires knowledge of the entire distribution, it is simple to simulate a sample with length T , to estimate the parameters, and to compute the KS test of the simulated standardized residuals (where the residuals are using estimated parameters). These steps can be repeated B times ($B > 1000$, possibly larger) and then the correct critical values can be computed from the empirical 90, 95 or 99% quantiles from KS_b , $b = 1, 2, \dots, B$. These quantiles are the correct values to use under the null while accounting for parameter estimation uncertainty.

8.7.6 Evaluating conditional density forecasts

In a direct analogue to the unconditional case, if $X_{t+1}|\mathcal{F}_t \sim F$, then $\hat{u}_{t+1} \equiv F(\hat{x}_{t+1})|\mathcal{F}_t \stackrel{\text{i.i.d.}}{\sim} U(0, 1)$. That is, the probability integral transformed residuals are *conditionally* i.i.d. Uniform(0, 1). While this condition is simple and easy to interpret, direct implementation of a test is not. The Berkowitz (2001) test works around this by further transforming the probability integral transformed residuals into normals using the inverse Normal CDF. Specifically if $\hat{u}_{t+1} = F_{t+1|t}(\hat{e}_{t+1})$ are the residuals standardized by their forecast distributions, the Berkowitz test computes $\hat{y}_{t+1} = \Phi^{-1}(\hat{u}_{t+1}) = \Phi^{-1}(F_{t+1|t}(\hat{e}_{t+1}))$ which have the property, under the null of a correct specification that $\hat{y}_t \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$ – an i.i.d. sequence of standard normal random variables.

Berkowitz proposes using a regression model to test the y_t for i.i.d. $N(0, 1)$. The test is implemented by estimating the parameters of

$$y_t = \phi_0 + \phi_1 y_{t-1} + \eta_t$$

via maximum likelihood. The Berkowitz test is computing using the likelihood ratio test

$$LR = 2(l(\hat{\theta}; \mathbf{y}) - l(\theta_0; \mathbf{y})) \sim \chi_3^2 \quad (8.40)$$

where θ_0 are the parameters if the null is true, $\phi_0 = \phi_1 = 0$ and $\sigma^2 = 1$ (3 restrictions). In other words, that the y_t are independent normal random variables with a variance of 1. As is always the case in tests of conditional models, the regression model can be augmented to include any time $t - 1$ available instrument and a more general specification is

$$y_t = \mathbf{x}_t \gamma + \eta_t$$

where \mathbf{x}_t may contain a constant, lagged y_t or anything else relevant for evaluating a density forecast. In the general specification, the null is $H_0 : \gamma = 0, \sigma^2 = 1$ and the alternative is the unrestricted estimate from the alternative specification. The likelihood ratio test statistic in the general case would have a χ_{K+1}^2 distribution where K is the number of elements in \mathbf{x}_t (the +1 comes from the restriction that $\sigma^2 = 1$).

8.8 Coherent Risk Measures

With multiple measures of risk available, which should be chosen: variance, VaR, or Expected Shortfall? Recent research into risk measurement has identified four desirable properties of any risk measure. Let ρ be any measure of risk, e.g., VaR or ES, that maps the riskiness of a portfolio to the reserves required to cover regularly occurring losses. P , P_1 and P_2 are portfolios of assets.

Drift Invariance

The required reserves for portfolio P satisfies

$$\rho(P + C) = \rho(P) - c$$

That is, adding a portfolio C with a constant return c to P decreases the required reserves by that amount.

Homogeneity

The required reserves are linear homogeneous,

$$\rho(\lambda P) = \lambda \rho(P) \quad \text{for any } \lambda > 0. \quad (8.41)$$

The homogeneity property states that the required reserves of two portfolios with the same relative holdings of assets depends linearly on the scale – doubling the size of a portfolio while not altering its relative composition generates twice the risk, and requires twice the reserves to cover regular losses.

Monotonicity

If P_1 first-order stochastically dominates P_2 (P_1 FOSD P_2), the required reserves for P_1 must be less than those of P_2 since

$$\rho(P_1) \leq \rho(P_2). \quad (8.42)$$

If P_1 FOSD P_2 then the value of portfolio P_1 is larger than the value of portfolio P_2 in every state of the world, and so the portfolio must be less risky.

Subadditivity

The required reserves for the combination of two portfolios is less than the required reserves for each treated separately

$$\rho(P_1 + P_2) \leq \rho(P_1) + \rho(P_2). \quad (8.43)$$

Definition 8.6 (Coherent Risk Measure). Any risk measure which satisfies these four properties is *coherent*.

Coherency seems like a good thing for a risk measure. The first three conditions are indisputable. For example, in the third, if P_1 FOSD P_2 , then P_1 always has a higher return, and so must be less risky. The last is somewhat controversial.

Theorem 8.2 (Value-at-Risk is not Coherent). *Value-at-Risk is not coherent since it fails the subadditivity criteria. It is possible to have a VaR which is superadditive where the Value-at-Risk of the combined portfolio is greater than the sum of the Values-at-Risk of either portfolio.*

Examples of the superadditivity of VaR usually require a portfolio for non-linear exposures. The simplest example where subadditivity fails is in portfolios of default bonds. Suppose P_1 and P_2 are portfolios where each contains a single bond with a face value of \$1,000 paying 0% interest. The bonds in the portfolios are from two companies. Assume that the default of one company is independent of the default of the other and that each company defaults with probability 3%. If a company defaults, only 60% of the bond value is recovered. The 5% VaR of both P_1 and P_2 is 0 since the companies pays the full \$1,000 97% of the time. The VaR of $P_3 = 50\% \times P_1 + 50\% \times P_2$, however, is \$200 since at least one company defaults 5.91% of the time. The distribution of P_3 is:

Probability	Portfolio Value
0.09%	\$600
5.82%	\$800
94.09%	\$1,000

Expected Shortfall, on the other hand, is a coherent measure of risk.

Theorem 8.3 (Expected Shortfall is Coherent). *Expected shortfall is a coherent risk measure.*

The proof that Expected Shortfall is coherent is straight forward in specific models (for example, if the returns are jointly normally distributed). The proof for an arbitrary distribution is challenging and provides little intuition. However, coherency alone does not make Expected Shortfall a better choice than VaR for measuring portfolio risk. VaR has many advantages as a risk measure: it only requires the modeling of a quantile of the return distribution, VaR always exists and is finite, and there are many well-established methodologies for accurately estimating VaR. Expected Shortfall requires an estimate of the mean in the tail which is more difficult to estimate accurately than the VaR. The ES may not exist in some cases if the distribution is very heavy-tailed. Additionally, in most realistic cases, increases in the Expected Shortfall is accompanied with increases in the VaR, and so these two measures often agree about the risk in a portfolio.

Shorter Problems

Problem 8.1. Discuss any properties the generalized error should have when evaluating Value-at-Risk models.

Problem 8.2. Define and contrast Historical Simulation and Filtered Historical Simulation?

Problem 8.3. Define Expected Shortfall. How does this extend the idea of Value-at-Risk? Why is it preferred to Value-at-Risk?

Problem 8.4. Why are HITs useful for testing a Value-at-Risk model?

Problem 8.5. Define conditional Value-at-Risk. Describe two methods for estimating this and compare their strengths and weaknesses.

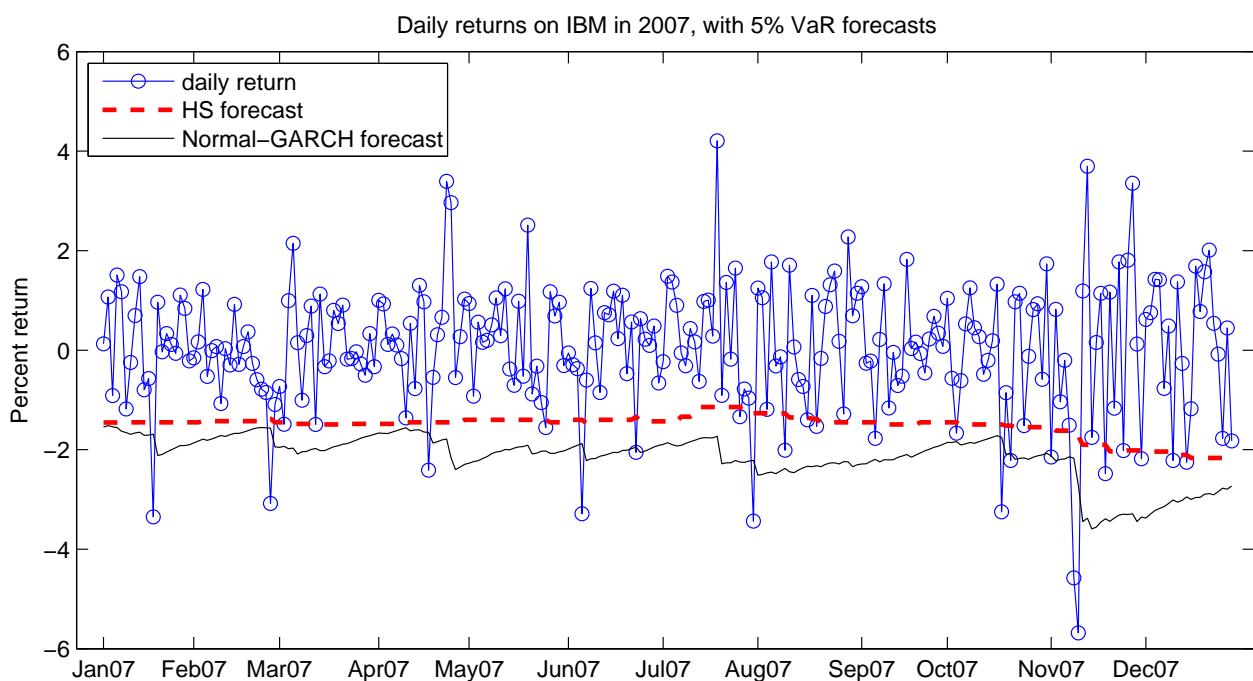
Problem 8.6. How are Value-at-Risk forecasts assessed? Describe two methods that can be sued to detect flawed Value-at-Risk models.

Longer Exercises

Exercise 8.1. Precisely answer the following questions.

1. What is VaR?
2. What is Expected Shortfall?
3. Describe two methods to estimate the VaR of a portfolio? Compare the strengths and weaknesses of these two approaches.
4. Suppose two bankers provide you with VaR forecasts (which are different) and you can get data on the actual portfolio returns. How could you test for superiority? What is meant by better forecast in this situation?

Exercise 8.2. The figure below plots the daily returns on IBM from 1 January 2007 to 31 December 2007 (251 trading days), along with 5% Value-at-Risk (VaR) forecasts from two models. The first model (denoted “HS”) uses Historical Simulation with a 250-day window of data. The second model uses a GARCH(1,1) model, assuming that daily returns have a constant conditional mean, and are conditionally Normally distributed (denoted “Normal-GARCH” in the figure).



1. Briefly describe **one** other model for VaR forecasting, and discuss its pros and cons relative to the Historical Simulation model and the Normal-GARCH model.

2. For each of the two VaR forecasts in the figure, a sequence of HIT variables was constructed:

$$\begin{aligned} HIT_t^{HS} &= \mathbf{1}\left\{r_t \leq \widehat{VaR}_t^{HS}\right\} \\ HIT_t^{GARCH} &= \mathbf{1}\left\{r_t \leq \widehat{VaR}_t^{GARCH}\right\} \\ \text{where } \mathbf{1}\{r_t \leq a\} &= \begin{cases} 1, & \text{if } r_t \leq a \\ 0, & \text{if } r_t > a \end{cases} \end{aligned}$$

and the following regression was run (standard errors are in parentheses below the parameter estimates):

$$\begin{aligned} HIT_t^{HS} &= 0.0956 + u_t \\ &\quad (0.0186) \\ HIT_t^{GARCH} &= 0.0438 + u_t \\ &\quad (0.0129) \end{aligned}$$

- (a) How can we use the above regression output to test the accuracy of the VaR forecasts from these two models?
- (b) What do the tests tell us?
3. Another set of regressions was also run (standard errors are in parentheses below the parameter estimates):

$$\begin{aligned} HIT_t^{HS} &= 0.1018 - 0.0601 HIT_{t-1}^{HS} + u_t \\ &\quad (0.0196) \quad (0.0634) \\ HIT_t^{GARCH} &= 0.0418 + 0.0491 HIT_{t-1}^{GARCH} + u_t \\ &\quad (0.0133) \quad (0.0634) \end{aligned}$$

A joint test that the intercept is 0.05 and the slope coefficient is zero yielded a chi-squared statistic of 6.9679 for the first regression, and 0.8113 for the second regression.

- (a) Why are these regressions potentially useful?
- (b) What do the results tell us? (The 95% critical values for a chi-squared variable with q degrees of freedom are given below:)

q	95% critical value
1	3.84
2	5.99
3	7.81
4	9.49
5	11.07
10	18.31
25	37.65
249	286.81
250	287.88
251	288.96

Exercise 8.3. Figure 8.10 plots the daily returns from 1 January 2008 to 31 December 2008 (252 trading days), along with 5% Value-at-Risk (VaR) forecasts from two models. The first model (denoted “HS”) uses *historical simulation* with a 250-day window of data. The second model uses a GARCH(1,1) model, assuming that daily returns have a constant conditional mean, and are conditionally Normally distributed (denoted “Normal-GARCH” in the figure).

1. Briefly describe **one** other model for VaR forecasting, and discuss its pros and cons relative to the Historical Simulation model and the Normal-GARCH model.
2. For each of the two VaR forecasts in the figure, a sequence of HIT variables was constructed:

$$\begin{aligned} HIT_t^{HS} &= \mathbf{1} \left\{ r_t \leq \widehat{VaR}_t^{HS} \right\} \\ HIT_t^{GARCH} &= \mathbf{1} \left\{ r_t \leq \widehat{VaR}_t^{GARCH} \right\} \\ \text{where } \mathbf{1} \{ r_t \leq a \} &= \begin{cases} 1, & \text{if } r_t \leq a \\ 0, & \text{if } r_t > a \end{cases} \end{aligned}$$

and the following regression was run (standard errors are in parentheses below the parameter estimates):

$$\begin{aligned} HIT_t^{HS} &= 0.0555 + u_t \\ &\quad (0.0144) \\ HIT_t^{GARCH} &= 0.0277 + u_t \\ &\quad (0.0103) \end{aligned}$$

- (a) How can we use the above regression output to test the accuracy of the VaR forecasts from these two models?
- (b) What do the tests tell us?
3. Another set of regressions was also run (standard errors are in parentheses below the parameter estimates):

$$\begin{aligned} HIT_t^{HS} &= 0.0462 + 0.1845 HIT_{t-1}^{HS} + u_t \\ &\quad (0.0136) \quad (0.1176) \\ HIT_t^{GARCH} &= 0.0285 - 0.0233 HIT_{t-1}^{GARCH} + u_t \\ &\quad (0.0106) \quad (0.0201) \end{aligned}$$

A joint test that the intercept is 0.05 and the slope coefficient is zero yielded a chi-squared statistic of 8.330 for the first regression, and 4.668 for the second regression.

- (a) Why are these regressions potentially useful?
- (b) What do the results tell us? (The 95% critical values for a chi-squared variable with q degrees of freedom are given below:)
4. Comment on the similarities and differences between what you found when testing using only a constant and when using a constant and the lagged HIT.

q	95% critical value
1	3.84
2	5.99
3	7.81
4	9.49
5	11.07
10	18.31
25	37.65
249	286.81
250	287.88
251	288.96

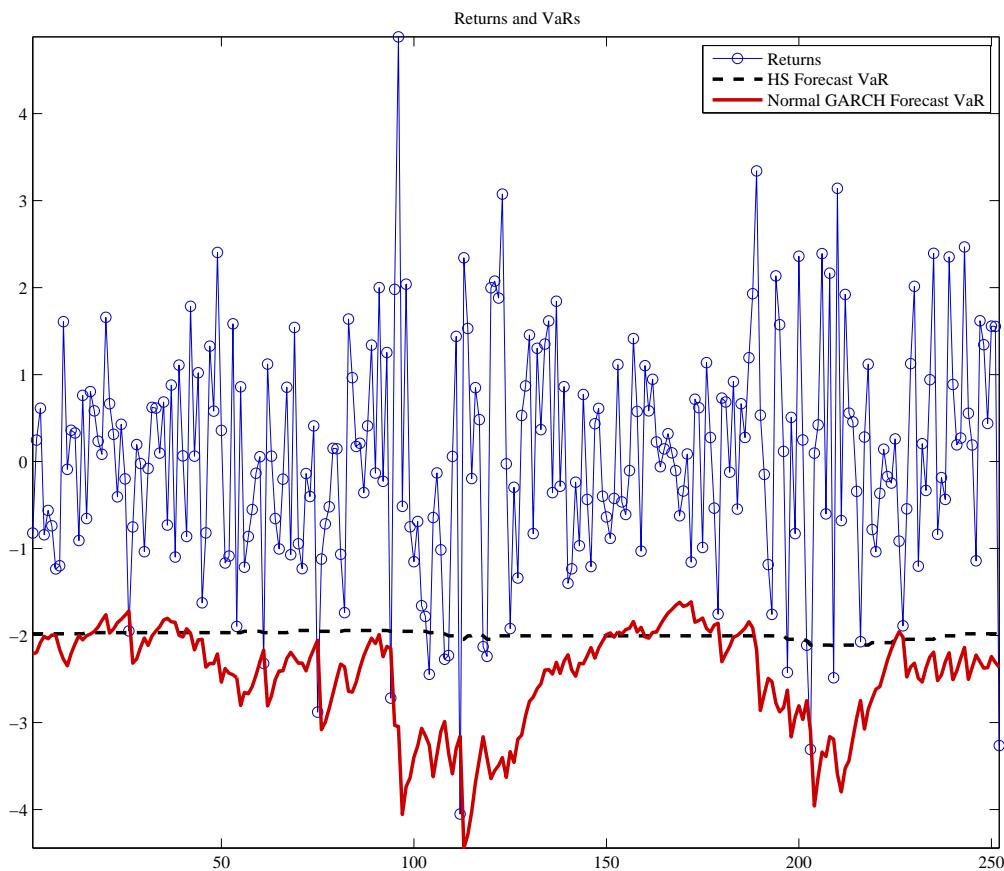


Figure 8.10: Returns, Historical Simulation VaR and Normal GARCH VaR.

Exercise 8.4. Answer the following question:

1. Assume that X is distributed according to some distribution F that is continuous and strictly increasing. Define $U \equiv F(X)$. Show that $U \sim \text{Uniform}(0, 1)$.
2. Assume that $V \sim \text{Uniform}(0, 1)$, and that G is some continuous and strictly increasing distribution function. If we define $Y \equiv G^{-1}(V)$, show that $Y \sim G$.

For the next two parts, consider the problem of forecasting the time taken for the price of a particular asset (P_t) to reach some threshold (P^*). Denote the time (in days) taken for the asset to reach the threshold as Z_t . Assume that the true distribution of Z_t is Exponential with parameter $\beta \in (0, \infty)$:

$$\begin{aligned} Z_t &\sim \text{Exponential}(\beta) \\ \text{so } F(z; \beta) &= \begin{cases} 1 - \exp\{-\beta z\}, & z \geq 0 \\ 0, & z < 0 \end{cases} \end{aligned}$$

Now consider a forecaster who gets the distribution correct, but the parameter wrong. Denote her distribution forecast as $\hat{F}(z) = \text{Exponential}(\hat{\beta})$.

3. If we define $U \equiv \hat{F}(Z)$, show that $\Pr[U \leq u] = 1 - (1 - u)^{\beta/\hat{\beta}}$ for $u \in (0, 1)$, and interpret.
4. Now think about the case where $\hat{\beta}$ is an estimate of β , such that $\hat{\beta} \xrightarrow{P} \beta$ as $n \rightarrow \infty$. Show that $\Pr[U \leq u] \xrightarrow{P} u$ as $n \rightarrow \infty$, and interpret.

Exercise 8.5. A Value-at-Risk model was fit to some return data, and the series of 5% VaR violations was computed. Denote these \widetilde{HIT}_t . The total number of observations was $T = 50$, and the total number of violations was 4.

1. Test the null that the model has unconditionally correct coverage using a t -test.
2. Test the null that the model has unconditionally correct coverage using a LR test. The likelihood for a Bernoulli(p) random Y is

$$f(y; p) = p^y (1 - p)^{1-y}.$$

The following regression was estimated

$$\widetilde{HIT}_t = 0.0205 + 0.7081 \widetilde{HIT}_{t-1} + \hat{\eta}_t$$

The estimated asymptotic covariance of the parameters is

$$\hat{\sigma}^2 \hat{\Sigma}_{XX}^{-1} = \begin{bmatrix} 0.0350 & -0.0350 \\ -0.0350 & 0.5001 \end{bmatrix}, \text{ and } \hat{\Sigma}_{XX}^{-1} \hat{\mathbf{S}} \hat{\Sigma}_{XX}^{-1} = \begin{bmatrix} 0.0216 & -0.0216 \\ -0.0216 & 2.8466 \end{bmatrix}$$

where $\hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^T \hat{\eta}_t^2$, $\hat{\Sigma}_{XX} = \frac{1}{T} \mathbf{X}' \mathbf{X}$ and $\hat{\mathbf{S}} = \frac{1}{T} \sum_{t=1}^T \hat{\eta}_t^2 \mathbf{x}_t' \mathbf{x}_t$.

3. Is there evidence that the model is dynamically misspecified, ignoring the unconditional rate of violations?
4. Compute a joint test that the model is completely correctly specified. Note that

$$\begin{bmatrix} a & b \\ b & c \end{bmatrix}^{-1} = \frac{1}{ac - b^2} \begin{bmatrix} c & -b \\ -b & a \end{bmatrix}.$$

Note: The 5% critical values of a χ^2_v are

v	CV
1	3.84
2	5.99
3	7.81
47	64.0
48	65.1
49	66.3
50	67.5

Exercise 8.6. Suppose you have a sample of 500 observations to evaluate a Value-at-Risk model using Out-of-Sample forecasts. You observe 36 95% VaR violations in this period.

1. What features the VaR violations of a correctly specified VaR model have?
2. Perform a test that the model is well specified using the sample average.
3. The likelihood of a Bernoulli(p) random variable is

$$L(y; p) = y^p (1-y)^{(1-p)}.$$

How can you use this likelihood to implement a better test? Compute the test statistic and draw conclusions about the accuracy of the model.

4. Explain the differences between these two approaches.
5. Fully describe one method that would allow you to use the time series of VaR violations to test whether the model has correctly specified dynamics.

Chapter 9

Multivariate Volatility, Dependence and Copulas

Modeling the conditional covariance of the assets in a portfolio is more challenging than modeling the variance of the portfolio. There are two challenges unique to the multivariate problem: ensuring that the conditional covariance is positive definite and finding a parsimonious specification that limits the number of model parameters in applications to large portfolios. This chapter covers standard moving-average covariance models, multivariate ARCH and Realized Covariance. While correlations are a key component of portfolio optimization, these measures are insufficient to fully characterize the joint behavior of asset returns, especially when markets are turbulent. This chapter introduces leading alternative measures of cross-asset dependence and then concludes with an introduction to a general framework for modeling multivariate returns using copulas.

9.1 Introduction

Multivariate volatility or covariance modeling is a crucial ingredient in modern portfolio management. It is applied to many important tasks, including:

- Portfolio Construction - Classic Markowitz (1959) portfolio construction requires an estimate of the covariance of returns, along with the expected returns of the assets, to determine the optimal portfolio weights. The Markowitz problem finds the portfolio with the minimum variance subject to achieving a required expected return. Alternatively, the Markowitz problem can be formulated as maximizing the expected mean of the portfolio given a constraint on the volatility of the portfolio.
- Portfolio Sensitivity Analysis - Many portfolios are constructed using objectives other than those in the Markowitz optimization problem. For example, fund managers may be selecting investment opportunities based on beliefs about fundamental imbalances between a firm and its competitors. Accurate measurement of asset return covariance is essential when assessing the portfolio's sensitivity to new positions. The sensitivity to the existing portfolio may be the deciding factor when evaluating multiple investment opportunities that have similar risk-return characteristics.

- Value-at-Risk - Naive $\alpha - VaR$ estimators scale the standard deviation of a portfolio by a constant value that depends on the quantile α . The conditional covariance allows the VaR sensitivity of the positions in the portfolio to be examined.
- Credit Pricing - Many credit products are written on a basket of bonds, and the correlation between the defaults of the underlying bonds is essential when determining the value of the derivative.
- Correlation Trading - Recent financial innovations allow correlation to be directly traded. The traded correlation is formally an equicorrelation (See 9.3.5). These products allow accurate correlation predictions to be used as the basis of a profitable trading strategy.

This chapter begins with an overview of simple, static estimators of covariance which are widely used. Attention then turns to dynamic models of conditional covariance based on the ARCH framework. Realized covariance, which exploits ultra-high frequency data in the same manner as realized variance, is then introduced as an improved estimator of the covariance. This chapter concludes with an examination of non-linear dependence measures and copulas, a recent introduction to financial econometrics that enables complex multivariate models to be flexibly constructed.

9.2 Preliminaries

Most volatility models are built using either returns, which is appropriate if the time horizon is small and the conditional mean is small relative to the conditional volatility or demeaned returns when using longer time-spans or if working with series with a non-trivial mean (e.g., electricity prices). The k by 1 vector of returns is denoted \mathbf{r}_t , and the demeaned returns are $\boldsymbol{\varepsilon}_t = \mathbf{r}_t - \boldsymbol{\mu}_t$ where $\boldsymbol{\mu}_t \equiv E_{t-1}[\mathbf{r}_t]$ is the conditional mean.

The conditional covariance, $\Sigma_t \equiv E_{t-1}[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}'_t]$, is assumed to be a k by k positive definite matrix. Some models make use of devolatilized residuals defined as $u_{i,t} = \boldsymbol{\varepsilon}_{i,t}/\sigma_{i,t}$, $i = 1, 2, \dots, k$, or in matrix notation $\mathbf{u}_t = \boldsymbol{\varepsilon}_t \oslash \boldsymbol{\sigma}_t$ where \oslash denotes Hadamard division (element-by-element) and $\boldsymbol{\sigma}_t$ is a k by vector of conditional standard deviations. Multivariate standardized residuals, which are both devolatilized and decorrelated, are defined $\mathbf{e}_t = \Sigma_t^{-\frac{1}{2}} \boldsymbol{\varepsilon}_t$ so that $E_{t-1}[\mathbf{e}_t \mathbf{e}'_t] = \mathbf{I}_k$. Some models explicitly parameterize the conditional correlation, $E_{t-1}[\mathbf{u}_t \mathbf{u}'_t] \equiv \mathbf{R}_t = \Sigma_t \oslash (\boldsymbol{\sigma}_t \boldsymbol{\sigma}'_t)$, or equivalently $\mathbf{R}_t = \mathbf{D}_t^{-1} \Sigma_t \mathbf{D}_t^{-1}$ where

$$\mathbf{D}_t = \begin{bmatrix} \sigma_{1,t} & 0 & \dots & 0 \\ 0 & \sigma_{2,t} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{k,t} \end{bmatrix}$$

and so $\Sigma_t = \mathbf{D}_t \mathbf{R}_t \mathbf{D}_t$.

Some models use a factor structure to reduce the dimension of the estimation problem. The p by 1 vector of factors is denoted \mathbf{f}_t and the factor returns are assumed to be mean $\mathbf{0}$, or demeaned if the assumption of conditional mean 0 is inappropriate. The conditional covariance of the factors is denoted $\Sigma_t^f \equiv E_{t-1}[\mathbf{f}_t \mathbf{f}'_t]$.

This chapter focuses exclusively on models capable of predicting the time- t covariance using information in \mathcal{F}_{t-1} . Multi-step forecasting is possible from many models in this chapter by direct

recursion, simulation or bootstrapping. Alternatively, direct forecasting techniques can be used to mix higher frequency data (e.g., daily) with longer forecast horizons (e.g., 2-week or one month).

9.2.1 Synchronization

Synchronization is a significant concern when measuring and modeling covariance, and non-synchronous returns can occur for a variety of reasons:

- Market opening and closing time differences – Most assets trade in liquid markets for only a fraction of the day. Differences in market hours frequently occur when modeling the return of assets that trade in different venues. The NYSE closes at either 20:00 or 21:00 GMT, depending on whether the U.S. east coast is using Eastern Standard or Daylight Time (EDT or EST). The London Stock Exchange closes at 15:30 or 16:30 GMT, depending on whether the U.K. is on British Summer Time (BST). Changes in U.S. equity prices that occur after the LSE closes are not reflected in U.K. equity prices until the next trading day.

Even within the same geographic region markets have different trading hours. Common U.S. equities trade from 9:30 until 16:00 EDT/EST time. U.S. government bond futures are traded using open outcry from 7:20 a.m. to 14:00. Light Sweet Crude futures trade 9:00 - 14:30 in an open outcry session. Closing prices, which are computed at the end of the trading day, do not reflect the same information in these three markets.

- Market closures due to public holidays – Markets are closed for public holidays which differ across geographies. Closures can even differ across markets, especially across asset class, within a country due to historical conventions.
- Delays in opening or closing – Assets that trade on the same exchange may be subject to opening or closing delays. For example, the gap between the first-to-open and the last-to-open stock in the S&P 500 can be as long as 15 minutes. While the range of closing times of the constituents is narrower, these are also not perfectly synchronized. These seemingly small differences lead to challenges when measuring the covariance using intra-daily (high-frequency) returns.
- Illiquidity/Stale Prices - Some assets trade more than others. The most liquid stock in the S&P 500 has a daily volume that is typically at least 100 times larger than the least liquid. Illiquidity is problematic when measuring covariance using intra-daily data.¹

There are three solutions to address biases that arise when modeling non-synchronous data. The first is to use relatively low-frequency returns. When using daily data, the NYSE and LSE are typically simultaneously open for 2 hours of 6½ hour U.S. trading day (30%). Using multi-day returns partially mitigates the lack of standardized opening hours since developments in U.S. equities on one day affect prices in London on the next day. For example, when using 2-day returns, it is as if 8.5 out of the 13 trading hours are synchronous (65%). When using weekly returns (5-day), 28 out of 32.5 hours are synchronized (86%). The downside of aggregating returns is the loss of data: parameter estimators

¹On February 26, 2010, the most liquid S&P 500 company was Bank of America (BAC) which had a volume of 96,352,600. The least liquid S&P 500 company was the Washington Post (WPO) which had a volume of 21,000. IMS Healthcare (RX) was acquired by another company, and so did not trade.

are less efficient when low-frequency return measurement makes it difficult to adjust portfolios for change in risk due to recent news.

The second solution is to use synchronized prices (also known as pseudo-closing prices). Synchronized prices are collected when all markets are simultaneously open. For example, if using prices of NYSE and LSE listed firms, returns constructed using prices sampled 1 hour before the LSE closes, which typically corresponds to 10:30 Eastern time, are synchronized. Daily returns constructed from these prices should capture all of the covariance between these assets. This approach is only a partial solution since many markets overlap in their trading hours, and so it is not applicable when measuring the covariance of a broad internationally diversified portfolio.

The third solution is to synchronize the non-synchronous returns using a vector moving average (Burns, Engle, and Mezrich, 1998). Suppose returns are ordered so that the first to close is in position 1, the second to close is in position 2, and so on until the last to close is in position k . With this ordering, returns on day $t+1$ for asset i may be correlated with the return on day t for asset j whenever $j > i$, and that the return on day $t+1$ should not be correlated with the day t return on asset j when $j \leq i$.

For example, consider modeling the leading equity index of the Australian Stock Exchange (UTC 0:00 - 6:10), the London Stock Exchange (UTC 8:00 - 16:30), NYSE (UTC 14:30 - 21:30) and Tokyo Stock Exchange (UTC 18:00 - 0:00 (+1 day)). The ASX closes before any of the others open. News from the ASX on day t appears in the LSE, NYSE, and TSE on the same day. The LSE opens second and so innovations in the LSE on day t may be correlated with changes on the ASX on $t+1$. Similarly, innovations in New York after UTC 16:30 affect $t+1$ returns in the ASX and LSE. Finally, news which comes out when the TSE is open shows up in the day $t+1$ return in the 3 other markets. This leads to a triangular structure in a vector moving average,

$$\begin{bmatrix} r_t^{\text{ASX}} \\ r_t^{\text{LSE}} \\ r_t^{\text{NYSE}} \\ r_t^{\text{TSE}} \end{bmatrix} = \begin{bmatrix} 0 & \theta_{12} & \theta_{13} & \theta_{14} \\ 0 & 0 & \theta_{23} & \theta_{24} \\ 0 & 0 & 0 & \theta_{34} \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \varepsilon_{t-1}^{\text{ASX}} \\ \varepsilon_{t-1}^{\text{LSE}} \\ \varepsilon_{t-1}^{\text{NYSE}} \\ \varepsilon_{t-1}^{\text{TSE}} \end{bmatrix} + \begin{bmatrix} \varepsilon_t^{\text{ASX}} \\ \varepsilon_t^{\text{LSE}} \\ \varepsilon_t^{\text{NYSE}} \\ \varepsilon_t^{\text{TSE}} \end{bmatrix} \quad (9.1)$$

The recursive structure of this system simplifies estimation since $r_t^{\text{TSE}} = \varepsilon_t^{\text{TSE}}$, and so the model for r_t^{NYSE} is a MA(1)-X. Given estimates of $\varepsilon_t^{\text{NYSE}}$, the model for r_t^{LSE} is also a MA(1)-X. This recursive MA(1)-X structure applies to the remaining assets in the model.

In vector form, this adjustment model is

$$\mathbf{r}_t = \Theta \varepsilon_{t-1} + \varepsilon_t$$

where \mathbf{r}_t is the k by 1 vector of nonsynchronous returns. Synchronized returns, $\hat{\mathbf{r}}_t$ are constructed using the VMA parameters as

$$\hat{\mathbf{r}}_t = (\mathbf{I}_k + \Theta) \varepsilon_t.$$

Θ captures any components in asset return j correlated with the return to asset return i when market i closes later than the where j . In essence this procedure “brings forward” the fraction of the return which has not yet occurred when asset j closes. Finally, the conditional covariance of ε_t is Σ_t , and so the covariance of the synchronized returns is $E_{t-1} [\hat{\mathbf{r}}_t \hat{\mathbf{r}}_t'] = (\mathbf{I}_k + \Theta) \Sigma_t (\mathbf{I}_k + \Theta)'$. Implementing this adjustment requires fitting the conditional covariance to the residual from the VMA, ε_t , rather than to returns directly.

9.3 Simple Models of Multivariate Volatility

Many simple models that rely on closed-form parameter estimators are widely used as benchmarks. These models are localized using rolling-windows, and so have a limited ability to adapt to changing market conditions.

9.3.1 Moving Average Covariance

The n -period moving average is the simplest covariance estimator.

Definition 9.1 (n -period Moving Average Covariance). The n -period moving average covariance is defined

$$\Sigma_t = n^{-1} \sum_{i=1}^n \boldsymbol{\varepsilon}_{t-i} \boldsymbol{\varepsilon}'_{t-i} \quad (9.2)$$

When returns are measured daily, standard choices for n are 22 (monthly), 66 (quarterly), or 252 (annual). When returns are measured monthly, standard choices for n are 12 (annual) or 60. When variance and correlations are time-varying, moving average covariances are imprecise measures; they simultaneously give too little weight to recent observations and place too much on observations in the distant past.

9.3.2 Exponentially Weighted Moving Average Covariance

Exponentially weighted moving averages (EWMA) are an alternative to moving average covariance estimators and allow for more weight on recent information. EWMA have been popularized in the volatility literature by RiskMetrics, which is introduced as a standard *VaR* model in chapter 8.

Definition 9.2 (Exponentially Weighted Moving Average Covariance). The EWMA covariance is defined recursively as

$$\Sigma_t = (1 - \lambda) \boldsymbol{\varepsilon}_{t-1} \boldsymbol{\varepsilon}'_{t-1} + \lambda \Sigma_{t-1} \quad (9.3)$$

for $\lambda \in (0, 1)$. EWMA covariance is equivalently defined through the infinite moving average

$$\Sigma_t = (1 - \lambda) \sum_{i=1}^{\infty} \lambda^{i-1} \boldsymbol{\varepsilon}_{t-i} \boldsymbol{\varepsilon}'_{t-i}. \quad (9.4)$$

An EWMA covariance estimator depends on an initial value for Σ_1 , which is usually set to the average covariance over the first m days for some $m > k$ or to the full-sample covariance. The single remaining parameter, λ , is usually to a value close to 1.

Definition 9.3 (RiskMetrics 1994 Covariance). The RiskMetrics 1994 Covariance is computed as an EWMA with $\lambda = .94$ for daily data or $\lambda = .99$ for monthly (J.P.Morgan/Reuters, 1996).

The RiskMetrics EWMA estimator, formally known as RM1994, has been updated to RM2006. The improved covariance estimator uses a model with a longer memory than RM1996. Long memory processes have weights that decay hyperbolically ($w \propto i^{-\alpha}$, $\alpha > 0$) rather than exponentially ($w \propto \lambda^i$). The new methodology extends the 1994 methodology by computing the volatility as a weighted sum of EWMA (eq. 9.5, line 1) rather than a single EWMA (eq. 9.3). This structure simplifies estimation since incorporating a new observation into the conditional covariance only requires updating the values of a small number of EWMA.

Definition 9.4 (RiskMetrics 2006 Covariance). The RiskMetrics 2006 Covariance is computed as

$$\begin{aligned}\Sigma_t &= \sum_{i=1}^m w_i \Sigma_{i,t} & (9.5) \\ \Sigma_{i,t} &= (1 - \lambda_i) \varepsilon_{t-1} \varepsilon'_{t-1} + \lambda_i \Sigma_{i,t-1} \\ w_i &= \frac{1}{C} (1 - \ln(\tau_i)/\ln(\tau_0)) \\ \lambda_i &= \exp(-1/\tau_i) \\ \tau_i &= \tau_1 \rho^{i-1}, \quad i = 1, 2, \dots, m\end{aligned}$$

where C is a constant that ensures that $\sum_{i=1}^m w_i = 1$.

The 2006 update is a 3-parameter model that includes a logarithmic decay factor, τ_0 (1560), a lower cut-off, τ_1 (4), and an upper cutoff τ_{\max} (512), where the suggested values are in parentheses. One additional parameter, ρ , is required to operationalize the model, and RiskMetrics suggests $\sqrt{2}$ (Zumbach, 2007).²

Both RiskMetrics covariance estimators can be expressed as weighted averages of the outer-products of shocks, $\Sigma_t = \sum_{i=1}^{\infty} \gamma_i \varepsilon_{t-i} \varepsilon'_{t-1}$, for a set of weights $\{\gamma_i\}$. Figure 9.1 contains a plot of the weights on the 120 most recent observations from both estimators. The updated methodology places both more weight on recent data and more weight on values in the distant past relative to the RM1996 model. Computing the number of periods before 99% of the weight is accumulated, or $\min_n \sum_{i=0}^n \gamma_i \geq .99$, is a simple method to compare the two methodologies. In RM1994, 99% of the weight accumulates in 75 observations when $\lambda = 0.94$ – the RM2006 methodology takes 619 days. The first 75 weights in the RM2006 model contain 83% of the weight, and so 1/6 of the total weight is assigned to returns more than 2 months in the past.

9.3.3 Observable Factor Covariance

The n -period factor model assumes that returns are generated by a strict factor structure and is closely related to the CAP-M (Sharpe, 1964; Lintner, 1965; Black, 1972), the intertemporal CAP-M (Merton, 1973) and Arbitrage Pricing Theory (Roll, 1977). Moving average factor covariance estimators are restricted moving average covariance estimators where the covariance between assets is attributed to common exposure to a set of factors. The model postulates that the return on the i^{th} asset is generated by a set of p observable factors with returns \mathbf{f}_t , an p by 1 set of asset-specific factor loadings, β_i and an idiosyncratic shock $\eta_{i,t}$,

$$\varepsilon_{i,t} = \mathbf{f}'_t \beta_{i,t} + \eta_{i,t}.$$

The k by 1 vector of returns is compactly described as

² τ_{\max} does not directly appear in the RM2006 framework, but is implicitly included since

$$m = 1 + \frac{\ln\left(\frac{\tau_{\max}}{\tau_1}\right)}{\ln \rho}.$$

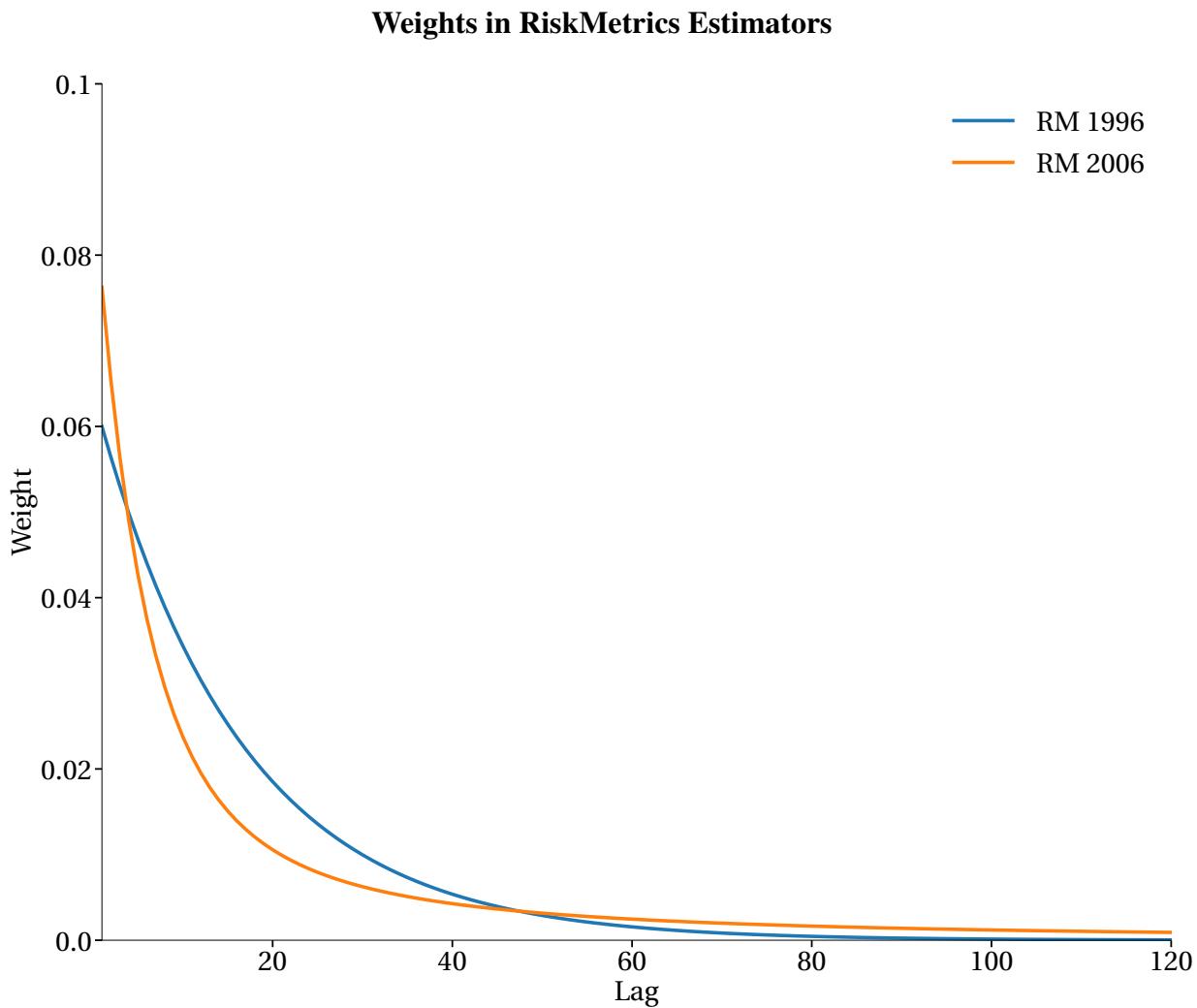


Figure 9.1: These two lines show the weights assigned to the lagged outer-product of returns ($\varepsilon_t \varepsilon_t'$) in the 1994 and 2006 versions of the RiskMetrics methodology. The 2006 version places more weight on recent shocks and more weight on shocks in the distant past relative to the 1994 methodology.

$$\varepsilon_t = \beta \mathbf{f}_t + \eta_t$$

where β is a k by p matrix of factor loadings and η_t is a k by 1 vector of idiosyncratic shocks. The shocks are assumed to be white noise, cross-sectionally uncorrelated ($E_{t-1} [\eta_{i,t} \eta_{j,t}] = 0$) and uncorrelated with the factors.

Definition 9.5 (n -period Factor Covariance). The n -period factor covariance is defined as

$$\Sigma_t = \beta \Sigma_t^f \beta' + \Omega_t \quad (9.6)$$

where $\Sigma_t^f = n^{-1} \sum_{i=1}^n \mathbf{f}_{t-i} \mathbf{f}'_{t-i}$ is the n -period moving covariance of the factors,

$$\beta_t = \left(\sum_{i=1}^n \mathbf{f}_{t-i} \mathbf{f}'_{t-i} \right)^{-1} \sum_{i=1}^n \mathbf{f}_{t-i} \varepsilon'_{t-i}$$

is the p by k matrix of factor loadings and Ω_t is a diagonal matrix with $\omega_{j,t}^2 = n^{-1} \sum_{i=1}^n \eta_{j,t-i}^2$ in the j^{th} diagonal position where $\eta_{i,t} = \varepsilon_{i,t} - \mathbf{f}'_t \beta_i$ are regression residuals.

Imposing a factor structure on the covariance has one key advantage: factor covariance estimators are positive definite when the number of periods used to estimate the factor covariance is larger than the number of factors ($n > p$). The standard moving average covariance estimator is only positive definite when the number of observations is larger than the number of assets ($n > k$). This feature facilitates application of factor covariance estimators in very large portfolios.

Structure can be imposed on the factor loadings estimator to improve covariance estimates in heterogeneous portfolios. Loadings on unrelated factors can be restricted to zero. For example, suppose a portfolio hold of equity and credit instruments, and that a total of 5 factors are used to model the covariance – one common to all assets, two specific to equities and two specific to bonds. The factor covariance is a 5 by 5 matrix, and the factor loadings for all assets have only three non-zero coefficients: the common factor and two asset-class specific factors. Zero restrictions on the factor loadings allow for application to large, complex portfolios, even in cases where many factors are needed to capture the systematic risk components in the portfolio.

9.3.4 Principal Component Covariance

Principal component analysis (PCA) is a statistical technique that decomposes a T by k matrix \mathbf{Y} into a T by k set of orthogonal (uncorrelated) factors, \mathbf{F} , and a k by k set of normalized weights (or factor loadings), β . Formally the principal component problem is defined as the solution

$$\arg \min_{\beta, \mathbf{F}} (kT)^{-1} \sum_{i=1}^k \sum_{t=1}^T (y_{i,t} - \mathbf{f}_t \beta_i)^2 \text{ subject to } \beta' \beta = \mathbf{I}_k \quad (9.7)$$

where \mathbf{f}_t is a 1 by k vector of common factors and β_i is a k by 1 vector of factor loadings. The solution to the principal component objective function can be computed from an eigenvalue decomposition of the outer product of \mathbf{Y} , $\mathbf{Y} = \mathbf{Y}' \mathbf{Y} = \sum_{t=1}^T \mathbf{y}_t \mathbf{y}'_t$.

Definition 9.6 (Orthonormal Matrix). A k -dimensional orthonormal matrix \mathbf{U} satisfies $\mathbf{U}' \mathbf{U} = \mathbf{I}_k$, and so $\mathbf{U}' = \mathbf{U}^{-1}$.

Definition 9.7 (Eigenvalues). The eigenvalues of a real, symmetric matrix k by k matrix \mathbf{A} are the k solutions to

$$|\lambda \mathbf{I}_k - \mathbf{A}| = 0 \quad (9.8)$$

where $|\cdot|$ is the determinant function.

Definition 9.8 (Eigenvectors). A k by 1 vector \mathbf{u} is an eigenvector corresponding to an eigenvalue λ of a real, symmetric matrix k by k matrix \mathbf{A} if

$$\mathbf{A}\mathbf{u} = \lambda \mathbf{u} \quad (9.9)$$

Theorem 9.1 (Spectral Decomposition Theorem). A real, symmetric matrix \mathbf{A} can be factored into $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}'$ where \mathbf{U} is an orthonormal matrix ($\mathbf{U}' = \mathbf{U}^{-1}$) containing the eigenvectors of \mathbf{A} in its columns and Λ is a diagonal matrix with the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_k$ of \mathbf{A} along its diagonal.

Since $\mathbf{Y}'\mathbf{Y} = \mathbf{\Upsilon}$ is real and symmetric with eigenvalues $\Lambda = \text{diag}(\lambda_i)_{i=1,\dots,k}$, the factors can be computed using the eigenvectors,

$$\begin{aligned}\mathbf{Y}'\mathbf{Y} &= \mathbf{U}\Lambda\mathbf{U}' \\ \mathbf{U}'\mathbf{Y}'\mathbf{Y}\mathbf{U} &= \mathbf{U}'\mathbf{U}\Lambda\mathbf{U}'\mathbf{U} \\ (\mathbf{Y}\mathbf{U})'(\mathbf{Y}\mathbf{U}) &= \Lambda && \text{since } \mathbf{U}' = \mathbf{U}^{-1} \\ \mathbf{F}'\mathbf{F} &= \Lambda.\end{aligned}$$

$\mathbf{F} = \mathbf{Y}\mathbf{U}$ is the T by k matrix of factors and $\beta = \mathbf{U}'$ is the k by k matrix of factor loadings. Additionally $\mathbf{F}\beta = \mathbf{F}\mathbf{U}' = \mathbf{Y}\mathbf{U}\mathbf{U}' = \mathbf{Y}$.³

The construction of the factor returns is the only difference between PCA-based covariance estimators and factor estimators. Factor estimators use observable portfolio returns to measure common exposure. In PCA-based covariance models, the factors are estimated from the returns, and so additional assets are not needed to measure common exposures.

Definition 9.9 (n -period Principal Component Covariance). The n -period principal component covariance is defined as

$$\Sigma_t = \beta_t'\Sigma_t^f\beta_t + \Omega_t \quad (9.10)$$

where $\Sigma_t^f = n^{-1} \sum_{i=1}^n \mathbf{f}_{t-i}\mathbf{f}_{t-i}'$ is the n -period moving covariance of first p principal component factors. $\hat{\beta}_t$ is the p by k matrix of principal component loadings corresponding to the first p factors. Ω_t is a diagonal matrix with diagonal elements $\omega_{j,t+1}^2 = n^{-1} \sum_{i=1}^n \eta_{i,t-1}^2$ where $\eta_{i,t} = r_{i,t} - \mathbf{f}_t'\beta_{i,t}$ are the residuals from a p -factor principal component analysis.

The number of factors, p , is the only parameter used to implement a PCA covariance estimator. The simple approach is to use a fixed number of factors based on experience or empirical regularities, e.g., selecting three factors when modeling with equity returns. The leading data-based approach is to select the number of factors by minimizing an information criterion such as those proposed in Bai and Ng (2002),

$$IC(p) = \ln(V(p, \hat{\mathbf{f}}^p)) + p \frac{k+T}{kT} \ln \left(\frac{kT}{k+T} \right)$$

where

$$V(p, \hat{\mathbf{f}}^p) = (kT)^{-1} \sum_{i=1}^k \sum_{t=1}^T \eta_{i,t}^2 \quad (9.11)$$

$$= (kT)^{-1} \sum_{i=1}^k \sum_{t=1}^T (r_{i,t} - \beta_i^p \mathbf{f}_t^p)^2 \quad (9.12)$$

³The factors and factor loadings are only identified up to ± 1 .

Principal Component Analysis of the S&P 500

$k = 378$	1	2	3	4	5	6	7	8	9	10
Partial R ²	0.327	0.038	0.035	0.025	0.023	0.018	0.015	0.010	0.010	0.008
Cumulative R ²	0.327	0.366	0.401	0.426	0.449	0.467	0.482	0.492	0.502	0.510

Table 9.1: Percentage of variance explained by the first 10 eigenvalues of the outer product matrix of S&P 500 returns. Returns on an asset are included if the asset is in the S&P 500 for 50% of the sample (k reports the number of firms that satisfy this criterion). The second line contains the cumulative R² of a p -factor model for the first 10 factors.

where β_i^p are the p factor loadings for asset i , and \mathbf{f}_t^p are the first p factors. The Bai and Ng information criterion is similar to other information criteria such as the HQIC or BIC. The first term, $\ln(V(p, \hat{\mathbf{f}}^p))$, measures the fit of a p -component model. Increasing p always improves the fit, and $p = \max(k, T)$ always perfectly explains the observed data. The second term, $p \frac{k+T}{kT} \ln(\frac{kT}{k+T})$, is a penalty that increases in p . Trading off these two leads to a consistent choice of p in data sets that are both long (large T) and wide (large k).

9.3.4.1 Interpreting the components

Factors extracted using PCA can be easily interpreted in terms of their contribution to total variance using R². This interpretation is possible since the factors are orthogonal, and so the R² of a model including $p < k$ factors is the sum of the R² of the p factors. Suppose the eigenvalues are ordered from largest to smallest and so $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k$ and that the factors associated with eigenvalue i are ordered such that it appears in column i of \mathbf{F} . The R² associated with factor i is then

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + \dots + \lambda_k},$$

and the cumulative R² of including $p < k$ factors is

$$\frac{\lambda_1 + \lambda_2 + \dots + \lambda_p}{\lambda_1 + \lambda_2 + \dots + \lambda_k}.$$

Cumulative R² is often used to select a subset of the k factors for model building. For example, in equity return data, it is not uncommon for 3–5 factors to explain 30-50% of the total variation in a large panel of equity returns.

9.3.4.2 Alternative methods

Principal components are often computed on either the covariance matrix of \mathbf{Y} or the correlation matrix of \mathbf{Y} . Using the covariance matrix is equivalent to building a model with an intercept,

$$y_{i,t} = \alpha_i + \mathbf{f}_t \beta_i \tag{9.13}$$

which differs from the principal components extracted from the outer product which is equivalent to the model

$$y_{i,t} = \mathbf{f}_t \boldsymbol{\beta}_i. \quad (9.14)$$

When working with asset return data, the difference between principal components extracted from the outer product and the covariance is negligible except in certain markets (e.g., electricity markets) or when using low-frequency returns (e.g., a month or more).

Principal components can also be extracted from the sample correlation matrix of \mathbf{Y} which is equivalent to the model

$$\frac{y_{i,t} - \bar{y}_i}{\hat{\sigma}_i} = \mathbf{f}_t \boldsymbol{\beta}_i \quad (9.15)$$

where $\bar{y}_i = T^{-1} \sum_{t=1}^T y_{i,t}$ is the mean of y_i and $\hat{\sigma}_i$ is the sample standard deviation of y_i . PCA is usually run on the correlation matrix when a subset of the series in \mathbf{Y} have variances which are much larger than the others. In cases where the variances differ greatly, principal components extracted from the outer product or covariance place more weight on the high variance series – fitting these high variance series produces the largest decrease in overall residual variance and the largest in R^2 for a fixed p . Using the correlation focuses the PCA estimator on the common (or systemic) variation rather than the variation of a small number of high variance asset returns.

9.3.5 Equicorrelation

Equicorrelation, like factor models, is a restricted covariance estimator. The equicorrelation estimator assumes that the covariance between any two assets can be expressed as $\rho \sigma_i \sigma_j$ where σ_i and σ_j are the volatilities of assets i and j , respectively. The correlation parameter is *not* indexed by i or j , and it is common to all assets. This estimator is misspecified whenever $k > 2$, and is generally only appropriate for assets where the majority of the pairwise correlations are homogeneous and positive.⁴

Definition 9.10 (n -period Moving Average Equicorrelation Covariance). The n -period moving average equicorrelation covariance is defined as

$$\Sigma_t = \begin{bmatrix} \sigma_{1,t}^2 & \rho_t \sigma_{1,t} \sigma_{2,t} & \rho_t \sigma_{1,t} \sigma_{3,t} & \dots & \rho_t \sigma_{1,t} \sigma_{k,t} \\ \rho_t \sigma_{1,t} \sigma_{2,t} & \sigma_{2,t}^2 & \rho_t \sigma_{2,t} \sigma_{3,t} & \dots & \rho_t \sigma_{2,t} \sigma_{k,t} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_t \sigma_{1,t} \sigma_{k,t} & \rho_t \sigma_{2,t} \sigma_{k,t} & \rho_t \sigma_{3,t} \sigma_{k,t} & \dots & \sigma_{k,t}^2 \end{bmatrix} \quad (9.16)$$

where $\sigma_{j,t}^2 = n^{-1} \sum_{i=1}^n \varepsilon_{j,t}^2$ and ρ_t is estimated using one of the estimators below.

The equicorrelation can be estimated using a moment-based estimator or a maximum-likelihood estimator. Define $\varepsilon_{p,t}$ as the equally weighted portfolio return. It is straightforward to see that

$$\begin{aligned} E[\varepsilon_{p,t}^2] &= k^{-2} \sum_{j=1}^k \sigma_{j,t}^2 + 2k^{-2} \sum_{o=1}^k \sum_{q=o+1}^k \rho \sigma_{o,t} \sigma_{q,t} \\ &= k^{-2} \sum_{j=1}^k \sigma_{j,t}^2 + 2\rho k^{-2} \sum_{o=1}^k \sum_{q=o+1}^k \sigma_{o,t} \sigma_{q,t} \end{aligned} \quad (9.17)$$

⁴The positivity constraint is needed to ensure that the covariance is positive definite which requires $\rho \in (-1/(k-1), 1)$, and so for k moderately large, the lower bound is effectively 0.

if the correlations among all of the pairs of assets are identical. The moment-based estimator replaces population values with estimates,

$$\sigma_{j,t}^2 = n^{-1} \sum_{i=1}^n \varepsilon_{j,t-i}^2, \quad j = 1, 2, \dots, k, p,$$

and the equicorrelation is estimated using

$$\rho_t = \frac{\sigma_{p,t}^2 - k^{-2} \sum_{j=1}^k \sigma_{j,t}^2}{2k^{-2} \sum_{o=1}^k \sum_{q=o+1}^k \sigma_{o,t} \sigma_{q,t}}.$$

Maximum likelihood, assuming returns are multivariate Gaussian, can alternatively be used to estimate the equicorrelation using standardized residuals $u_{j,t} = \varepsilon_{j,t} / \sigma_{j,t}$. The estimator for ρ can be found by maximizing the likelihood

$$\begin{aligned} L(\rho_t; \mathbf{u}) &= -\frac{1}{2} \sum_{i=1}^n k \ln 2\pi + \ln |\mathbf{R}_t| + \mathbf{u}'_{t-i} \mathbf{R}_t^{-1} \mathbf{u}_{t-i} \\ &= \sum_{i=1}^n k \ln 2\pi + \ln \left((1 - \rho_t)^{k-1} (1 + (k-1)\rho_t) \right) \\ &\quad + \frac{1}{(1 - \rho_t)} \left[\sum_{j=1}^k u_{j,t-i}^2 - \frac{\rho_t}{1 + (k-1)\rho_t} \left(\sum_{q=1}^k u_{q,t-i} \right)^2 \right] \end{aligned} \tag{9.18}$$

where \mathbf{u}_t is a k by 1 vector of standardized residuals and \mathbf{R}_t is a correlation matrix with all non-diagonal elements equal to ρ . This likelihood is computationally similar to univariate likelihood for any k and so maximization is very fast even when k is large.⁵

9.3.6 Application: S&P 500

The S&P 500 is used to illustrate the moving-average covariance estimators. The CRSP database provides daily return data for all constituents of the S&P 500. The sample runs from January 1, 1984, until December 31, 2018. The returns on firms are included in the data set is available for at least 50% of the sample.⁶

Table 9.1 contains the number of assets which meet this criterion (k) and both the partial and cumulative R² for the first 10 principal components. The first explains a substantial amount of the data (32.7%) and the next four combine to explain 42.6% of the cross-sectional variation. If returns did not follow a factor structure, then each principal component is expected to explain approximately 0.25% of the variation. Table 9.2 contains the full-sample equicorrelation, 1-factor R² using the S&P

⁵The computation speed of the likelihood can be increased by pre-computing $\sum_{j=1}^k u_{j,t-i}^2$ and $\sum_{q=1}^k u_{q,t-i}$.

⁶The Expectations-Maximization algorithm allows PCA to be applied in data sets containing missing values. The algorithm begins with a guess for the missing values, usually the mean of the non-missing values for each variable. The augmented data set is then used to estimate a p factor model (the *maximization* step). The missing values are then replaced with the fitted p components (the *expectations* step). These two steps are repeated until the process converges in the sense that the change in the fitted values for the missing coefficients is small.

Correlation Measures for the S&P 500

$k = 378$	Equicorrelation	1-Factor R^2 (S&P 500)	3-Factor R^2 (Fama-French)
	0.291	0.282	0.313

Table 9.2: Full sample correlation measures of the S&P 500. Returns on an asset are included if the asset is in the S&P 500 for more than 50% of the sample (k reports the number of firms that satisfy this criterion). The 1-factor R^2 is from a model using the return on the S&P 500, and the 3-factor R^2 is from a model that uses the returns on the 3 Fama-French portfolios.

500 index as the observable factor and the 3-factor R^2 using the 3 Fama-French portfolios as factors.⁷ The average correlation and the 1-factor fit is similar to that in the 1-factor PCA model, although the 3 Fama-French factors do not appear to work as well as the 3 factors estimated from the data. The difference between the 1- and 3-factor observable and PCA models is due to the lack of cross-sectional variation in firm size among the components of the S&P 500 when compared to all assets in CRSP.

Figure 9.2 contains a plot of the 1-year moving average equicorrelation and 1- and 3-factor PCA R^2 . Each component asset is included in the calculation if all returns are present in the 1-year window. Periods of high volatility, such as the end of the dot-com bubble and late 2008, also have a high correlation. The three lines broadly agree about the changes and only differ in level. Figure 9.3 contains plots of the R^2 from the 1-factor PCA and the 1-factor model which uses the S&P500 return (top panel) and the 3-factor PCA and the 3 Fama-French factors (bottom panel). The dynamics in all series are similar, and only the levels differ. PCA selects the factors to maximizes the fit in the cross-section, and so must produce a higher R^2 than the observable models for a given number of factors.

9.4 Multivariate ARCH Models

9.4.1 Vector GARCH (*vec*)

The Vector GARCH model uses a specification that naturally extends the univariate GARCH model to a model of the conditional covariance (Bollerslev, Engle, and Wooldridge, 1988). The model is defined using the *vec* of the conditional covariance, which stacks the elements of the covariance into a vector.

Definition 9.11 (Vector GARCH). The covariance in a vector GARCH(1,1) model (*vec*) evolves according to

$$\text{vec}(\Sigma_t) = \text{vec}(\mathbf{C}) + \mathbf{A}\text{vec}(\boldsymbol{\varepsilon}_{t-1}\boldsymbol{\varepsilon}'_{t-1}) + \mathbf{B}\text{vec}(\Sigma_{t-1}) \quad (9.19)$$

$$= \text{vec}(\mathbf{C}) + \mathbf{A}\text{vec}\left(\Sigma_{t-1}^{1/2}\mathbf{e}_t\left(\Sigma_{t-1}^{1/2}\mathbf{e}_t\right)'\right) + \mathbf{B}\text{vec}(\Sigma_{t-1}) \quad (9.20)$$

⁷These estimators are computed using missing values. The observable factor models are estimated using only the common sample where the factor and the individual asset are present. The equicorrelation is estimated by standardizing each series to have mean 0 and unit variance, and then computing the MLE of the correlation of these values treated as if they are a single series.

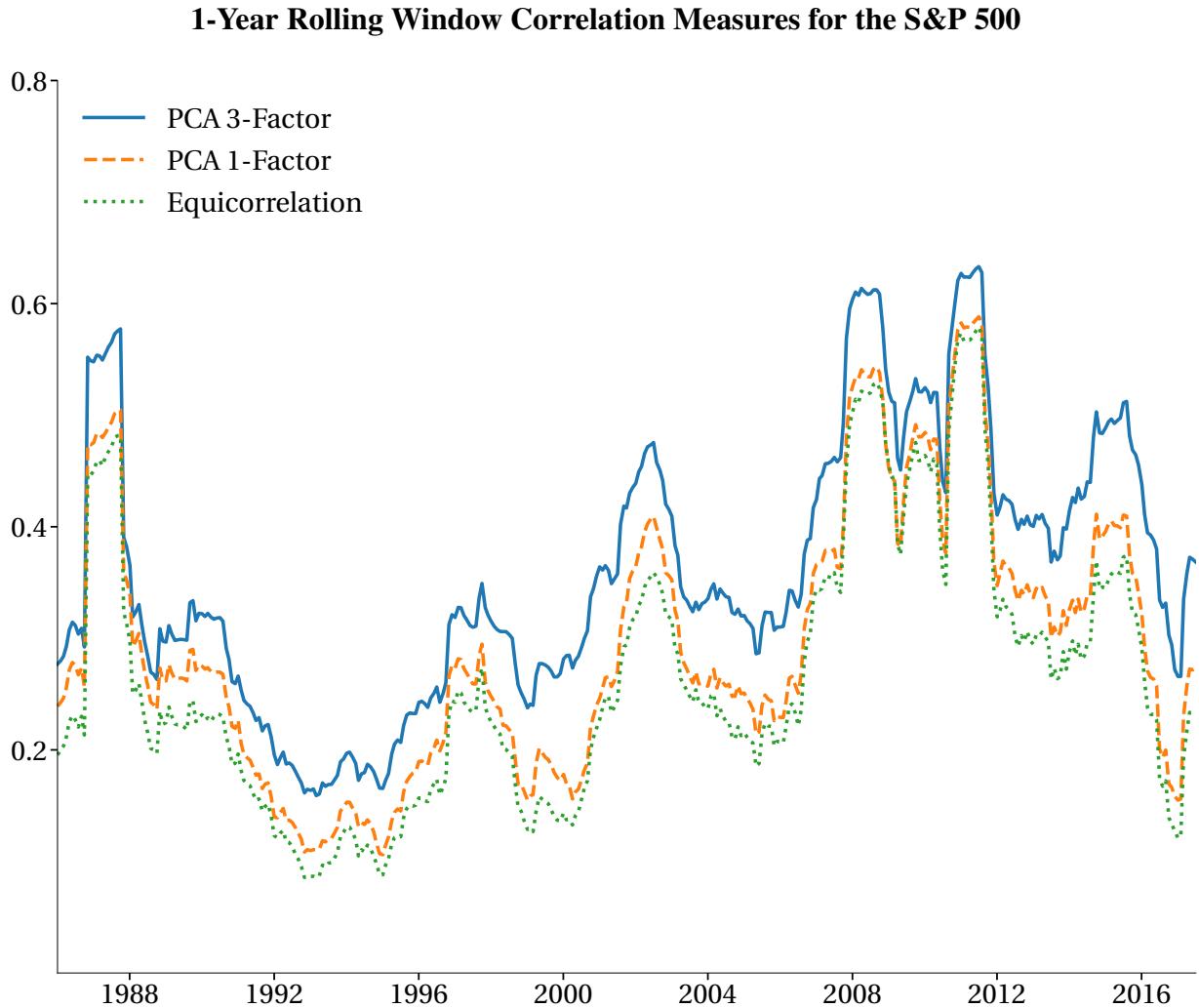
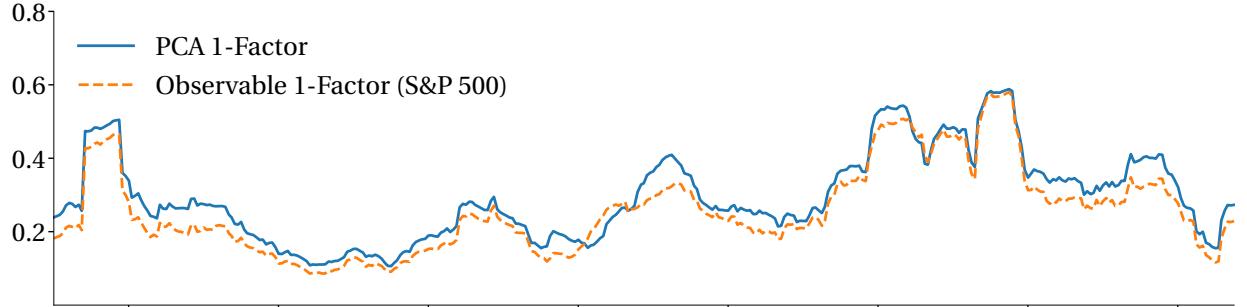


Figure 9.2: Three views of the average correlation of the S&P 500. The PCA measures are the R^2 of models with 1 and 3 factors. Each estimate is computed using a 1-year rolling window and is plotted against the center of the rolling window. All three measures roughly agree about the changes in the average correlation.

where \mathbf{C} is a k by k positive definite matrix and both \mathbf{A} and \mathbf{B} are k^2 by k^2 parameter matrices. $\Sigma_{t-1}^{1/2}$ is a matrix square root and $\{\mathbf{e}_t\}$ is a sequence of i.i.d. random variables with mean 0 and covariance \mathbf{I}_k , such as a standard multivariate normal.

See eq. 5.9 for the definition of the vec operator. The vec allows each square or cross-product to influence each term in the conditional covariance. To understand the richness of the specification, consider the evolution of the conditional covariance in a bivariate model,

Observable and Principal Component Correlation Measures for the S&P 500
1-Factor Models



3-Factor Models

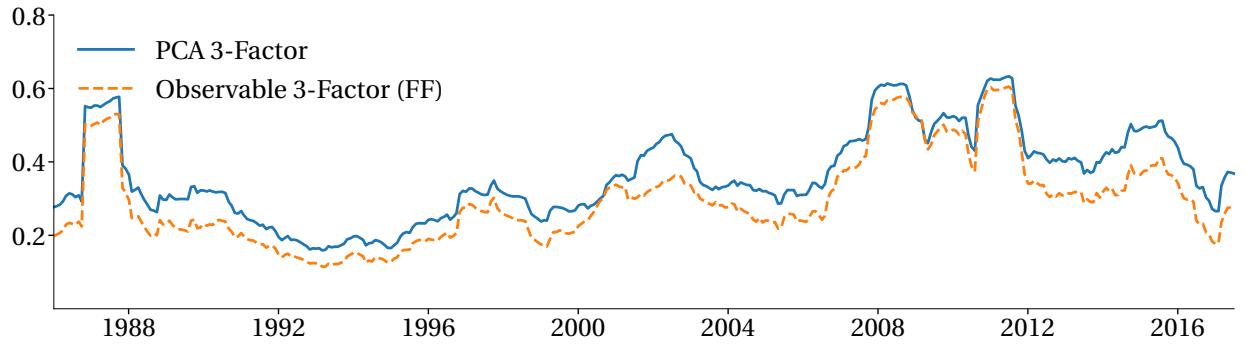


Figure 9.3: The top panel plots the R^2 for 1-factor PCA and an observable factor model which uses the return on the S&P 500 as the observable factor. The bottom contains the same for 3-factor PCA and the Fama-French 3-factor model. Each estimate is computed using a 1-year rolling window and is plotted against the center of the rolling window.

$$\begin{bmatrix} \sigma_{11,t} \\ \sigma_{12,t} \\ \sigma_{12,t} \\ \sigma_{22,t} \end{bmatrix} = \begin{bmatrix} c_{11} \\ c_{12} \\ c_{12} \\ c_{22} \end{bmatrix} + \begin{bmatrix} a_{11} & a_{12} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{22} & a_{23} \\ a_{21} & a_{22} & a_{22} & a_{22} \\ a_{41} & a_{42} & a_{42} & a_{44} \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t-1}^2 \\ \varepsilon_{1,t-1}\varepsilon_{2,t-1} \\ \varepsilon_{1,t-1}\varepsilon_{2,t-1} \\ \varepsilon_{2,t-1}^2 \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{22} & b_{23} \\ b_{21} & b_{22} & b_{22} & b_{23} \\ b_{41} & b_{42} & b_{42} & b_{44} \end{bmatrix} \begin{bmatrix} \sigma_{11,t-1} \\ \sigma_{12,t-1} \\ \sigma_{12,t-1} \\ \sigma_{22,t-1} \end{bmatrix}.$$

The repeated elements are needed to ensure that the conditional covariance is symmetric.

The vec operator stacks the elements of the covariance matrix and the outer products of returns. The evolution of the conditional variance of the first asset,

$$\sigma_{11,t} = c_{11} + a_{11}\varepsilon_{1,t-1}^2 + 2a_{12}\varepsilon_{1,t-1}\varepsilon_{2,t-1} + a_{13}\varepsilon_{2,t-1}^2 + b_{11}\sigma_{11,t-1} + 2b_{12}\sigma_{12,t-1} + b_{13}\sigma_{22,t-1},$$

depends on both past squared returns and the cross-product. In practice, it is difficult to use the vector GARCH model since it is challenging to determine the restrictions on \mathbf{A} and \mathbf{B} necessary to guarantee that Σ_t is positive definite.

The diagonal vec model has been more successful, primarily because it is relatively straight forward to find conditions which ensure that the conditional covariance is positive semi-definite. The diagonal vec model restricts \mathbf{A} and \mathbf{B} to be diagonal matrices so that the elements of Σ_t evolve according to

$$\Sigma_t = \mathbf{C} + \tilde{\mathbf{A}} \odot \varepsilon_{t-1} \varepsilon'_{t-1} + \tilde{\mathbf{B}} \odot \Sigma_{t-1} \quad (9.21)$$

where $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$ are symmetric parameter matrices and \odot the is Hadamard product operator.⁸ All elements of Σ_t evolve using GARCH(1,1)-like dynamics, so that

$$\sigma_{ij,t} = c_{ij} + \tilde{a}_{ij} \varepsilon_{i,t-1} \varepsilon_{j,t-1} + \tilde{b}_{ij} \sigma_{ij,t-1}.$$

The diagonal vec still requires restrictions on the parameters to ensure that the conditional covariance is positive definite. Ding and Engle (2001) develop one set of sufficient constraints on the parameters in the Matrix GARCH model (see section 9.4.3).

9.4.2 BEKK GARCH

The BEKK (Baba, Engle, Kraft, and Kroner) GARCH model directly addresses the difficulties in determining constraints on the parameters in a vec specification (Engle and Kroner, 1995). BEKK models rely on two results from linear algebra to ensure that the conditional covariance is positive definite: quadratic forms are positive semi-definite, and the sum of a positive semi-definite matrix and a positive definite matrix is positive definite.

Definition 9.14 (BEKK GARCH). The covariance in a BEKK GARCH(1,1) model evolves according to

$$\Sigma_t = \mathbf{CC}' + \mathbf{A}\varepsilon_{t-1}\varepsilon'_{t-1}\mathbf{A}' + \mathbf{B}\Sigma_{t-1}\mathbf{B}' \quad (9.22)$$

where \mathbf{C} is a k by k lower triangular matrix and \mathbf{A} and \mathbf{B} are k by k parameter matrices.

The BEKK is a restricted version of the vec specification where $\mathbf{A} \otimes \mathbf{A}$ and $\mathbf{B} \otimes \mathbf{B}$ control the response to recent news and the smoothing, respectively,

$$\text{vec}(\Sigma_t) = \text{vec}(\mathbf{CC}') + \mathbf{A} \otimes \mathbf{A} \text{vec}(\varepsilon_{t-1}\varepsilon'_{t-1}) + \mathbf{B} \otimes \mathbf{B} \text{vec}(\Sigma_{t-1}). \quad (9.23)$$

The elements of Σ_t depend on all cross-products. For example, in a bivariate BEKK,

8

Definition 9.12 (Hadamard Product). Let \mathbf{A} and \mathbf{B} be matrices with the same size. The Hadamard product of \mathbf{A} and \mathbf{B} , denoted $\mathbf{A} \odot \mathbf{B}$, is the matrix with ij^{th} element $a_{ij}b_{ij}$.

Definition 9.13 (Hadamard Quotient). Let \mathbf{A} and \mathbf{B} be matrices with the same size. The Hadamard quotient of \mathbf{A} and \mathbf{B} , denoted $\mathbf{A} \oslash \mathbf{B}$, is the matrix with ij^{th} element a_{ij}/b_{ij} .

$$\begin{bmatrix} \sigma_{11,t} & \sigma_{12,t} \\ \sigma_{12,t} & \sigma_{22,t} \end{bmatrix} = \begin{bmatrix} c_{11} & 0 \\ c_{12} & c_{22} \end{bmatrix} \begin{bmatrix} c_{11} & 0 \\ c_{21} & c_{22} \end{bmatrix}' + \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t-1}^2 & \varepsilon_{1,t-1}\varepsilon_{2,t-1} \\ \varepsilon_{1,t-1}\varepsilon_{2,t-1} & \varepsilon_{2,t-1}^2 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}' + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} \sigma_{11,t-1} & \sigma_{12,t-1} \\ \sigma_{12,t-1} & \sigma_{22,t-1} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}'. \quad (9.24)$$

The conditional variance of the first asset is

$$\sigma_{11,t} = c_{11}^2 + a_{11}^2 \varepsilon_{1,t-1}^2 + 2a_{11}a_{12}\varepsilon_{1,t-1}\varepsilon_{2,t-1} + a_{12}^2 \varepsilon_{2,t-1}^2 + b_{11}^2 \sigma_{11,t-1} + 2b_{11}b_{12}\sigma_{12,t-1} + b_{12}^2 \sigma_{22,t-1}.$$

The other conditional variance and the conditional covariance have similar forms that depend on both squared returns and the cross-product. Estimation of full BEKK models is difficult in portfolios with only a moderate number of assets since as the number of parameters in the model is $(5k^2 + k)/2$, and so is usually only appropriate for $k \leq 5$.

The diagonal BEKK partially addresses the growth rate in the number of parameters by restricting \mathbf{A} and \mathbf{B} to be diagonal matrices,

Definition 9.15 (Diagonal BEKK GARCH). The covariance in a diagonal BEKK GARCH(1,1) model evolves according to

$$\Sigma_t = \mathbf{C}\mathbf{C}' + \tilde{\mathbf{A}}\varepsilon_{t-1}\varepsilon_{t-1}'\tilde{\mathbf{A}}' + \tilde{\mathbf{B}}\Sigma_{t-1}\tilde{\mathbf{B}}'. \quad (9.25)$$

where \mathbf{C} is a k by k lower triangular matrix and $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$ are diagonal parameter matrices.

The conditional covariances in a diagonal BEKK evolve according to

$$\sigma_{ij,t} = \tilde{c}_{ij} + a_i a_j \varepsilon_{i,t-1} \varepsilon_{j,t-1} + b_i b_j \sigma_{ij,t-1} \quad (9.26)$$

where \tilde{c}_{ij} is the ij^{th} element of $\mathbf{C}\mathbf{C}'$. This specification is similar to the diagonal vec except that the parameters are shared across series.

The scalar BEKK further restricts the parameter matrices to be common across all assets and is a particularly simple (and restrictive) model.

Definition 9.16 (Scalar BEKK GARCH). The covariance in a scalar BEKK GARCH(1,1) model evolves according to

$$\Sigma_t = \mathbf{C}\mathbf{C}' + a^2 \varepsilon_{t-1} \varepsilon_{t-1}' + b^2 \Sigma_{t-1} \quad (9.27)$$

where \mathbf{C} is a k by k lower triangular matrix and a and b are scalar parameters.

The scalar BEKK has one further advantage: it can be covariance targeted, which simplifies parameter estimation. Covariance targeting replaces the intercept ($\mathbf{C}\mathbf{C}'$) with a consistent estimator, $(1 - a^2 - b^2)\bar{\Sigma}$, where $\bar{\Sigma}$ is the long-run covariance, $E[\Sigma_t]$. $\bar{\Sigma}$ is estimated using the outer product of returns, $\hat{\bar{\Sigma}} = T^{-1} \sum_{t=1}^T \varepsilon_t \varepsilon_t'$. The two remaining parameters, a and b , are then estimated conditioning on the estimate of the unconditional covariance of returns,

$$\Sigma_t = (1 - a^2 - b^2) \widehat{\Sigma} + a^2 \varepsilon_{t-1} \varepsilon'_{t-1} + b^2 \Sigma_{t-1}. \quad (9.28)$$

This 2-step estimator reduces the number of parameters that need to be simultaneously estimated using numerical methods from $2 + k(k + 1)/2$ to 2. The reduction in the parameter space allows covariance-targeted scalar BEKK models to be applied in large portfolios ($k > 50$).

9.4.3 Matrix GARCH (M-GARCH)

Matrix GARCH imposes structure on the parameters of a diagonal vec that ensure that the estimated conditional covariances are positive definite (Ding and Engle, 2001).

Definition 9.17 (Matrix GARCH). The covariance in a Matrix GARCH(1,1) model evolves according to

$$\Sigma_t = \mathbf{C}\mathbf{C}' + \mathbf{A}\mathbf{A}' \odot \varepsilon_{t-1} \varepsilon'_{t-1} + \mathbf{B}\mathbf{B}' \odot \Sigma_{t-1} \quad (9.29)$$

where \mathbf{C} , \mathbf{A} and \mathbf{B} are lower triangular matrices.

Ding and Engle (2001) show that if \mathbf{U} and \mathbf{V} are positive semi-definite matrices, then $\mathbf{U} \odot \mathbf{V}$ is also. Combining this result with the result that quadratic forms are positive semi-definite ensures that Σ_t is positive definite if \mathbf{C} has full rank. The diagonal Matrix GARCH, which restricts \mathbf{A} and \mathbf{B} to be vectors, is equivalent to the diagonal BEKK model.

Definition 9.18 (Diagonal Matrix GARCH). The covariance in a diagonal Matrix GARCH(1,1) model evolves according to

$$\Sigma_t = \mathbf{C}\mathbf{C}' + \mathbf{a}\mathbf{a}' \odot \varepsilon_{t-1} \varepsilon'_{t-1} + \mathbf{b}\mathbf{b}' \odot \Sigma_{t-1} \quad (9.30)$$

where \mathbf{C} is a lower triangular matrix and \mathbf{a} and \mathbf{b} are k by 1 parameter vectors. The scalar version of the Matrix GARCH is identical to the scalar BEKK.

9.4.4 Constant Conditional Correlation (CCC) GARCH

Constant Conditional Correlation GARCH Bollerslev (1990) uses a different approach to modeling the conditional covariance. CCC GARCH decomposes the conditional covariance into k conditional variances and the conditional correlation, which is assumed to be constant,

$$\Sigma_t = \mathbf{D}_t \mathbf{R} \mathbf{D}_t. \quad (9.31)$$

\mathbf{D}_t is a diagonal matrix composed of the conditional standard deviations,

$$\mathbf{D}_t = \begin{bmatrix} \sigma_{1,t} & 0 & 0 & \dots & 0 \\ 0 & \sigma_{2,t} & 0 & \dots & 0 \\ 0 & 0 & \sigma_{3,t} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \sigma_{k,t} \end{bmatrix} \quad (9.32)$$

where $\sigma_{i,t} = \sqrt{\sigma_{ii,t}}$ is the standard deviation of the i^{th} asset return. The conditional variances are typically modeled using GARCH(1,1) models,

$$\sigma_{ii,t} = \omega_i + \alpha_i \varepsilon_{i,t-1}^2 + \beta_i \sigma_{ii,t-1} \quad (9.33)$$

where $u_{i,t-1}$ is the i^{th} element of $\mathbf{u}_t = \mathbf{R}^{1/2} \mathbf{e}_t$ and $\{\mathbf{e}_t\}$ is a sequence of i.i.d. random variables with mean 0 and covariance \mathbf{I}_k . Other specifications, such as TARCH or EGARCH, can also be used to model the conditional variance. It is even possible to model the conditional variances using different models for each asset, which is a distinct advantage over *vec* and related models which impose common structure. The conditional correlation is constant

$$\mathbf{R} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \dots & \rho_{1k} \\ \rho_{12} & 1 & \rho_{23} & \dots & \rho_{2k} \\ \rho_{13} & \rho_{23} & 1 & \dots & \rho_{3k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{1k} & \rho_{2k} & \rho_{3k} & \dots & 1 \end{bmatrix}. \quad (9.34)$$

The conditional covariance matrix is constructed from the conditional standard deviations and the conditional correlation, and so all of the dynamics in the conditional covariance are attributable to changes in the conditional variances.

$$\Sigma_t = \begin{bmatrix} \sigma_{11,t} & \rho_{12}\sigma_{1,t}\sigma_{2,t} & \rho_{13}\sigma_{1,t}\sigma_{3,t} & \dots & \rho_{1k}\sigma_{1,t}\sigma_{k,t} \\ \rho_{12}\sigma_{1,t}\sigma_{2,t} & \sigma_{22,t} & \rho_{23}\sigma_{2,t}\sigma_{3,t} & \dots & \rho_{2k}\sigma_{2,t}\sigma_{k,t} \\ \rho_{13}\sigma_{1,t}\sigma_{3,t} & \rho_{23}\sigma_{2,t}\sigma_{3,t} & \sigma_{33,t} & \dots & \rho_{3k}\sigma_{3,t}\sigma_{k,t} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{1k}\sigma_{1,t}\sigma_{k,t} & \rho_{2k}\sigma_{2,t}\sigma_{k,t} & \rho_{3k}\sigma_{3,t}\sigma_{k,t} & \dots & \sigma_{kk,t} \end{bmatrix}. \quad (9.35)$$

Bollerslev (1990) shows that the CCC GARCH model can be estimated in two steps. The first fits k conditional variance models (e.g., GARCH) and produces the vector of standardized residuals \mathbf{u}_t where $u_{i,t} = \varepsilon_{i,t} / \sqrt{\hat{\sigma}_{ii,t}}$. The second step estimates the constant conditional correlation using the standard correlation estimator applied to the standardized residuals.

Definition 9.19 (Constant Conditional Correlation GARCH). The covariance in a constant conditional correlation GARCH model evolves according to

$$\Sigma_t = \begin{bmatrix} \sigma_{11,t} & \rho_{12}\sigma_{1,t}\sigma_{2,t} & \rho_{13}\sigma_{1,t}\sigma_{3,t} & \dots & \rho_{1k}\sigma_{1,t}\sigma_{k,t} \\ \rho_{12}\sigma_{1,t}\sigma_{2,t} & \sigma_{22,t} & \rho_{23}\sigma_{2,t}\sigma_{3,t} & \dots & \rho_{2k}\sigma_{2,t}\sigma_{k,t} \\ \rho_{13}\sigma_{1,t}\sigma_{3,t} & \rho_{23}\sigma_{2,t}\sigma_{3,t} & \sigma_{33,t} & \dots & \rho_{3k}\sigma_{3,t}\sigma_{k,t} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{1k}\sigma_{1,t}\sigma_{k,t} & \rho_{2k}\sigma_{2,t}\sigma_{k,t} & \rho_{3k}\sigma_{3,t}\sigma_{k,t} & \dots & \sigma_{kk,t} \end{bmatrix} \quad (9.36)$$

where $\sigma_{ii,t}, i = 1, 2, \dots, k$ evolve according to some univariate GARCH process, usually a GARCH(1,1).

9.4.5 Dynamic Conditional Correlation (DCC)

Dynamic Conditional Correlation extends CCC GARCH by introducing scalar BEKK-like dynamics to the conditional correlations, and so \mathbf{R} in the CCC is replaced with \mathbf{R}_t in the DCC (Engle, 2002b)

Definition 9.20 (Dynamic Conditional Correlation GARCH). The covariance in a Dynamic Conditional Correlation (DCC)-GARCH model evolves according to

$$\Sigma_t = \mathbf{D}_t \mathbf{R}_t \mathbf{D}_t. \quad (9.37)$$

where

$$\mathbf{R}_t = \mathbf{Q}_t^* \mathbf{Q}_t \mathbf{Q}_t^*, \quad (9.38)$$

$$\mathbf{Q}_t = (1 - a - b) \bar{\mathbf{R}} + a \mathbf{u}_{t-1} \mathbf{u}'_{t-1} + b \mathbf{Q}_{t-1}, \quad (9.39)$$

$$= (1 - a - b) \bar{\mathbf{R}} + a \left(\mathbf{R}_{t-1}^{\frac{1}{2}} \mathbf{e}_{t-1} \right) \left(\mathbf{R}_{t-1}^{\frac{1}{2}} \mathbf{e}_{t-1} \right)' + b \mathbf{Q}_{t-1}, \quad (9.40)$$

$$\mathbf{Q}_t^* = (\mathbf{Q}_t \odot \mathbf{I}_k)^{-\frac{1}{2}} \quad (9.41)$$

\mathbf{u}_t is the k by 1 vector of standardized returns ($u_{i,t} = \varepsilon_{i,t} / \sqrt{\hat{\sigma}_{ii,t}}$) and \odot denotes Hadamard multiplication (element-by-element). $\{\mathbf{e}_t\}$ are a sequence of i.i.d. innovations with mean $\mathbf{0}$ and covariance \mathbf{I}_k . \mathbf{D}_t is a diagonal matrix with the conditional standard deviation of asset i on the i^{th} diagonal position. The conditional variance, $\sigma_{ii,t}$, $i = 1, 2, \dots, k$, evolve according to some univariate GARCH process for asset i , usually a GARCH(1,1) and are identical to eq. 9.33.

The \mathbf{Q}_t process resembles a covariance targeting BEKK (eq. 9.28). Eqs. 9.38 and 9.41 are needed to ensure that \mathbf{R}_t is a correlation matrix with diagonal elements equal to 1. This structure allows for three-step estimation. The first two steps are identical to those in the CCC GARCH model. The third step conditions on the estimate of the long-run correlation when estimating the parameters of the dynamics, a and b .⁹

9.4.6 Orthogonal GARCH (OGARCH)

The principal components of a T by k matrix of returns ε are defined as $\mathbf{F} = \varepsilon \mathbf{U}$ where \mathbf{U} is the matrix of eigenvectors of the outer product of ε . Orthogonal GARCH uses the first p principal components to model the conditional covariance by assuming that the factors are conditionally uncorrelated.¹⁰

Definition 9.21 (Orthogonal GARCH). The covariance in an orthogonal GARCH (OGARCH) model evolves according to

$$\Sigma_t = \beta \Sigma_t^f \beta' + \Omega \quad (9.42)$$

where β is the k by p matrix of factor loadings corresponding to the p factors with the highest total R^2 . The conditional covariance of the factors is assumed diagonal,

$$\Sigma_t^f = \begin{bmatrix} \psi_{1,t}^2 & 0 & 0 & \dots & 0 \\ 0 & \psi_{2,t}^2 & 0 & \dots & 0 \\ 0 & 0 & \psi_{3,t}^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \psi_{l,t}^2 \end{bmatrix}, \quad (9.43)$$

⁹The three-step estimator is biased, even in large samples. Only two-step estimation – where the variances are first estimated, and then all correlation parameters are jointly estimated – produces consistent parameter estimates in DCC models.

¹⁰Principal components are estimated using the outer-product or the unconditional covariance of returns, and so only guarantee that the factors are unconditionally uncorrelated.

and the conditional variance of each factor follows a GARCH(1,1) process (other models possible)

$$\psi_{i,t}^2 = \varphi_i + \alpha_i f_{i,t-1}^2 + \beta_i \psi_{i,t-1}^2 \quad (9.44)$$

$$= \varphi_i + \alpha_i \psi_{i,t-1}^2 e_{t,t-1}^2 + \beta_i \psi_{i,t-1}^2 \quad (9.45)$$

where $\{\mathbf{e}_t\}$ is a sequence of i.i.d. innovations with mean $\mathbf{0}$ and covariance \mathbf{I}_k .

The conditional covariance of the residuals is assumed to be constant and diagonal,

$$\Omega = \begin{bmatrix} \omega_1^2 & 0 & 0 & \dots & 0 \\ 0 & \omega_2^2 & 0 & \dots & 0 \\ 0 & 0 & \omega_3^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \omega_l^2 \end{bmatrix}, \quad (9.46)$$

where each variance is estimated using the residuals from the p factor model,

$$\omega_i^2 = \sum_{t=1}^T \eta_{i,t}^2 = \sum_{t=1}^T (\varepsilon_{i,t} - \mathbf{f}_t \boldsymbol{\beta}_i)^2. \quad (9.47)$$

Variants of the standard OGARCH model include parameterizations where the number of factors is equal to the number of assets, and so $\Omega = \mathbf{0}$, and a specification which replaces Ω with Ω_t where each $\omega_{i,t}^2$ follows a univariate GARCH process.

9.4.7 Conditional Asymmetries

Standard multivariate ARCH models are symmetric since they only depend on the outer product of returns, and so have news impact curves that are identical for ε_t and $-\varepsilon_t$. Most models can be modified to allow for conditional asymmetries in covariance, a feature that may be important when modeling returns in some asset classes, e.g., equities. Define $\zeta_t = \varepsilon_t \odot I_{[\varepsilon_t < 0]}$ where $I_{[\varepsilon_t < 0]}$ is a k by 1 vector of indicator variables where the i^{th} position is 1 if $r_{i,t} < 0$. An asymmetric BEKK model can be constructed as

$$\Sigma_t = \mathbf{C}\mathbf{C}' + \mathbf{A}\varepsilon_{t-1}\varepsilon'_{t-1}\mathbf{A}' + \mathbf{G}\zeta_{t-1}\zeta'_{t-1}\mathbf{G}' + \mathbf{B}\Sigma_{t-1}\mathbf{B}' \quad (9.48)$$

where \mathbf{G} is a k by k matrix of parameters that measure the sensitivity to “bad” news. When $k = 1$, this model reduces to a GJR-GARCH(1,1,1) model for the variance. Diagonal and scalar BEKK models can be similarly adapted.

An asymmetric version of Matrix GARCH can be similarly constructed so that

$$\Sigma_t = \mathbf{C}\mathbf{C}' + \mathbf{A}\mathbf{A}' \odot \varepsilon_{t-1}\varepsilon'_{t-1} + \mathbf{G}\mathbf{G}' \odot \zeta_{t-1}\zeta'_{t-1} + \mathbf{B}\mathbf{B}' \odot \Sigma_{t-1} \quad (9.49)$$

where \mathbf{G} is a lower triangular parameter matrix. The dynamics of the covariances in the asymmetric Matrix GARCH process are

$$\sigma_{ij,t} = \tilde{c}_{ij} + \tilde{a}_{ij}r_{i,t-1}r_{j,t-1} + \tilde{g}_{ij}r_{i,t-1}r_{j,t-1}I_{i,t-1}I_{j,t-1} + \tilde{b}_{ij}\sigma_{ij,-t1}$$

where \tilde{c}_{ij} is the ij^{th} element of $\mathbf{C}\mathbf{C}'$ and \tilde{a}_{ij} , \tilde{g}_{ij} and \tilde{b}_{ij} are similarly defined. All conditional variances follow GJR-GARCH(1,1,1) models, and covariances evolve using similar dynamics driven by cross

products of returns. The asymmetry only has an effect on the conditional covariance between two assets if both markets experience “bad” news (negative returns). Cappiello, Engle, and Sheppard (2006) propose an asymmetric extension to the DCC model.

9.4.8 Fitting Multivariate GARCH Models

Returns are typically assumed to be conditionally multivariate normal, and so model parameters are estimated by maximizing the corresponding likelihood function,

$$f(\varepsilon_t; \theta) = (2\pi)^{-\frac{k}{2}} |\Sigma_t|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \varepsilon_t' \Sigma_t^{-1} \varepsilon_t\right) \quad (9.50)$$

where θ contains the collection of parameters in the model. Estimation is, in principle, a simple problem. In practice, parameter estimation is only straight-forward when the number of assets is relatively small (less than 10) or when the model is tightly parameterized (e.g., scalar BEKK). The log-likelihood in larger, more complex models is difficult to optimize for two reasons. First, the likelihood is relatively flat and so finding its maximum value is difficult for optimization software. Second, the computational cost of evaluating the log-likelihood is increasing in the number of unknown parameters and grows at rate k^3 in most multivariate ARCH models.

Many models have been designed to use multi-stage estimation to avoid these problems, including:

- *Covariance Targeting BEKK*: The intercept is concentrated out using the sample covariance of returns, and so only the parameters governing the dynamics of the conditional covariance need to be estimated using numerical methods.
- *Constant Conditional Correlation*: Fitting a CCC GARCH involves fitting k univariate GARCH models and then using a closed-form estimator of the constant conditional correlation.
- *Dynamic Conditional Correlation*: Fitting a DCC GARCH combines the first stage of the CCC GARCH with correlation targeting similar to that in covariance targeting BEKK.
- *Orthogonal GARCH*: Orthogonal GARCH only involves fitting $p \leq k$ univariate GARCH models and uses a closed-form estimator of the idiosyncratic variance.

9.4.9 Application: Mutual Fund Returns

Three mutual funds are used to illustrate the differences (and similarities) of multivariate ARCH models. The three funds are:

- Oakmark I (OAKMX), a large-cap fund;
- Fidelity Small Cap Stock (FSLCX), a small-cap fund which seeks to invest in firms with capitalizations similar to those in the Russell 2000 or S&P 600; and
- Wasatch-Hoisington US Treasury (WHOSX), a fund which invests at least 90% of AUM in U.S. Treasury securities.

CCC GARCH Correlation			
	Large Cap	Small Cap	Bond
Large Cap	1	0.718	-0.258
Small Cap	0.718	1	-0.259
Bond	-0.258	-0.259	1

Unconditional Correlation			
	Large Cap	Small Cap	Bond
Large Cap	1	0.803	-0.306
Small Cap	0.803	1	-0.305
Bond	-0.306	-0.305	1

Table 9.3: The top panel reports the estimates of the conditional correlation from a CCC GARCH model for three mutual funds spanning large-cap stocks (OAKMX), small-cap stocks (FSLCX), and long government bond returns (WHOSX). The bottom panel contains the estimtes of the unconditional correlation computed from the unfiltered returns.

All data comes from the CRSP database, and data between January 1, 1998, and December 31, 2018, is used to estimate model parameters. Table 9.3 contains the estimated correlation from the CCC-GARCH model where each volatility series is modeled using a GARCH(1,1). The correlations between these assets are large and positive for the equity funds and negative, on average, between the equity funds and the bond fund. The bottom panel reports the unconditional correlation of the returns. These values are all larger in magnitude than the conditional correlations. The conditional volatilities of the three series tend to comove, and the periods with high volatility have a disproportionate impact on the estimated covariance.

Table 9.4 contains the parameters of the dynamics of six models: the DCC, scalar BEKK, an asymmetric scalar BEKK, Matrix GARCH and the asymmetric extension of the Matrix GARCH model. The estimates of the parameters in the DCC are typical – the two parameters sum to nearly 1 and α is smaller in magnitude than the values typically found in volatility models. These estimates indicate that correlation is very persistent but less dynamic than volatility. The parameters in the scalar BEKK and asymmetric scalar BEKK are similar to what one typically finds in a volatility model, although the asymmetry is weak. The Matrix GARCH parameters are fairly homogeneous although the treasury fund is less responsive to news (i.e., has smaller coefficient in \mathbf{AA}'). In the asymmetric Matrix GARCH model, the response to “bad” news is not homogeneous. The equities have large asymmetries while the bond fund does not. This heterogeneity explains the small asymmetry parameter in the asymmetric scalar BEKK.

Figure 9.4 plots the annualized volatility for these series from four models: the CCC (standard GARCH(1,1)), the two RiskMetrics methodologies, and the asymmetric scalar BEKK. All volatilities are similar which is surprising given the differences in the models. Figures 9.5, 9.6 and 9.7 plot the correlations as fit from 6 different models. Aside from the correlation estimated in the CCC GARCH (which is constant), the estimated correlations are also substantially similar.

Multivariate GARCH Model Estimates									
	α		γ		β				
DCC		0.009 (3.4)		–	0.990 (4.9)				
Scalar BEKK		0.062 (143.0)		–	0.918 (89.6)				
Asym. Scalar BEKK	0.056 (158.9)	0.021 (84.7)	0.911 (65.8)						

	AA'			GG'			BB'		
Matrix GARCH	0.092 (5.37)	0.090 (6.03)	0.048 (2.34)	–	–	–	0.875 (36.74)	0.885 (32.35)	0.910 (14.62)
	0.090 (6.03)	0.087 (5.06)	0.047 (2.85)	–	–	–	0.885 (32.35)	0.895 (30.99)	0.921 (20.78)
	0.048 (2.34)	0.047 (2.85)	0.043 (6.20)	–	–	–	0.910 (14.62)	0.921 (20.78)	0.947 (111.86)
Asymmetric Matrix GARCH	0.073 (2.86)	0.068 (4.33)	0.050 (2.51)	0.038 (0.99)	0.042 (1.35)	–0.007 (−0.31)	0.872 (30.58)	0.883 (23.59)	0.908 (15.76)
	0.068 (4.33)	0.063 (1.44)	0.048 (2.68)	0.042 (1.35)	0.047 (1.50)	–0.008 (−0.44)	0.883 (23.59)	0.893 (17.55)	0.919 (19.18)
	0.050 (2.51)	0.048 (2.68)	0.043 (6.38)	–0.007 (−0.31)	–0.008 (−0.44)	0.001 (0.08)	0.908 (15.76)	0.919 (19.18)	0.946 (69.10)

Table 9.4: Parameter estimates (t -stats in parenthesis) from multivariate ARCH models for three mutual funds representing distinct investment styles: small-cap stocks (FSLCX), large-cap stocks (OAKMX), and long government bond returns (WHOSX). The top panel contains results for DCC, scalar BEKK and asymmetric scalar BEKK. The bottom panel contains estimation results for Matrix GARCH and the asymmetric extension to the Matrix GARCH model.

9.5 Realized Covariance

Realized Covariance uses ultra-high-frequency data (trade data) to estimate the integrated covariance over some period, usually a day. Suppose prices followed a k -variate continuous time diffusion,

$$d\mathbf{p}_t = \mu_t dt + \Omega_t d\mathbf{W}_t$$

where μ_t is the instantaneous drift, $\Sigma_t = \Omega_t \Omega_t'$ is the instantaneous covariance, and $d\mathbf{W}_t$ is a k -variate Brownian motion. Realized Covariance estimates

$$\int_0^1 \Sigma_s ds$$

where the bounds 0 and 1 represent the (arbitrary) interval over which the covariance is estimated. The integrated covariance is the multivariate analog of the integrated variance introduced in Chapter 7.¹¹

Realized covariance is computed using the outer-product of high-frequency returns.

¹¹In the presence of jumps, Realized Covariance estimates the quadratic covariation, which is the integrated covariance

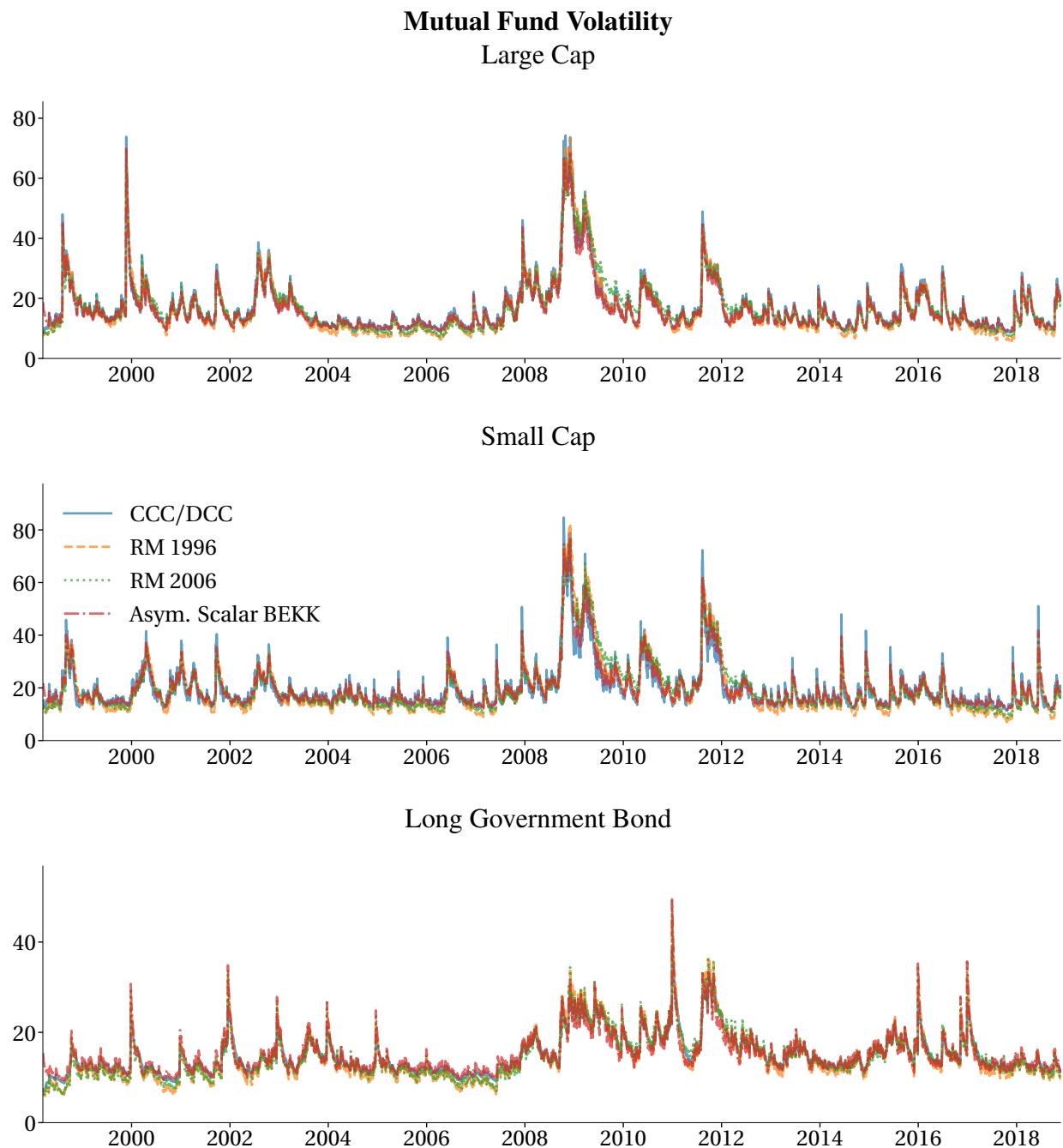


Figure 9.4: The three panels plot the estimated annualized volatility of the three mutual funds.

plus the outer product of the jumps

$$\int_0^1 \Sigma_s ds + \sum_{0 \leq s \leq 1} \Delta p_s \Delta p'_s,$$

where Δp_s are the jumps.

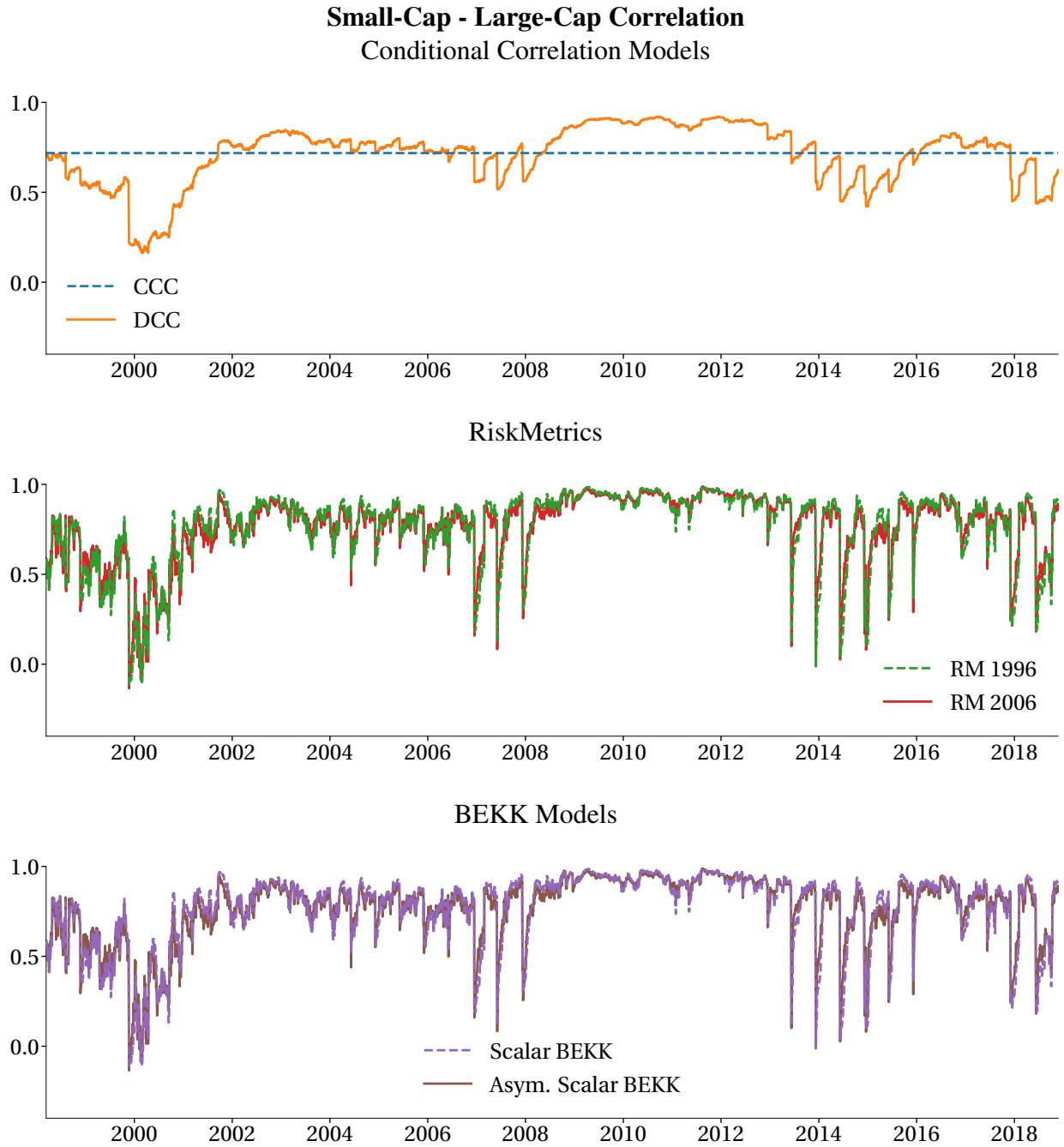


Figure 9.5: The three panels show the estimated conditional correlation between the large-cap fund and the small-cap fund from 6 models.

Definition 9.22 (Realized Covariance). The m -sample Realized Covariance is defined

$$RC_t^{(m)} = \sum_{i=1}^m \mathbf{r}_{i,t} \mathbf{r}'_{i,t} = (\mathbf{p}_{i,t} - \mathbf{p}_{i-1,t}) (\mathbf{p}_{i,t} - \mathbf{p}_{i-1,t})', \quad (9.51)$$

where $\mathbf{r}_{i,t}$ is the i^{th} return on day t .

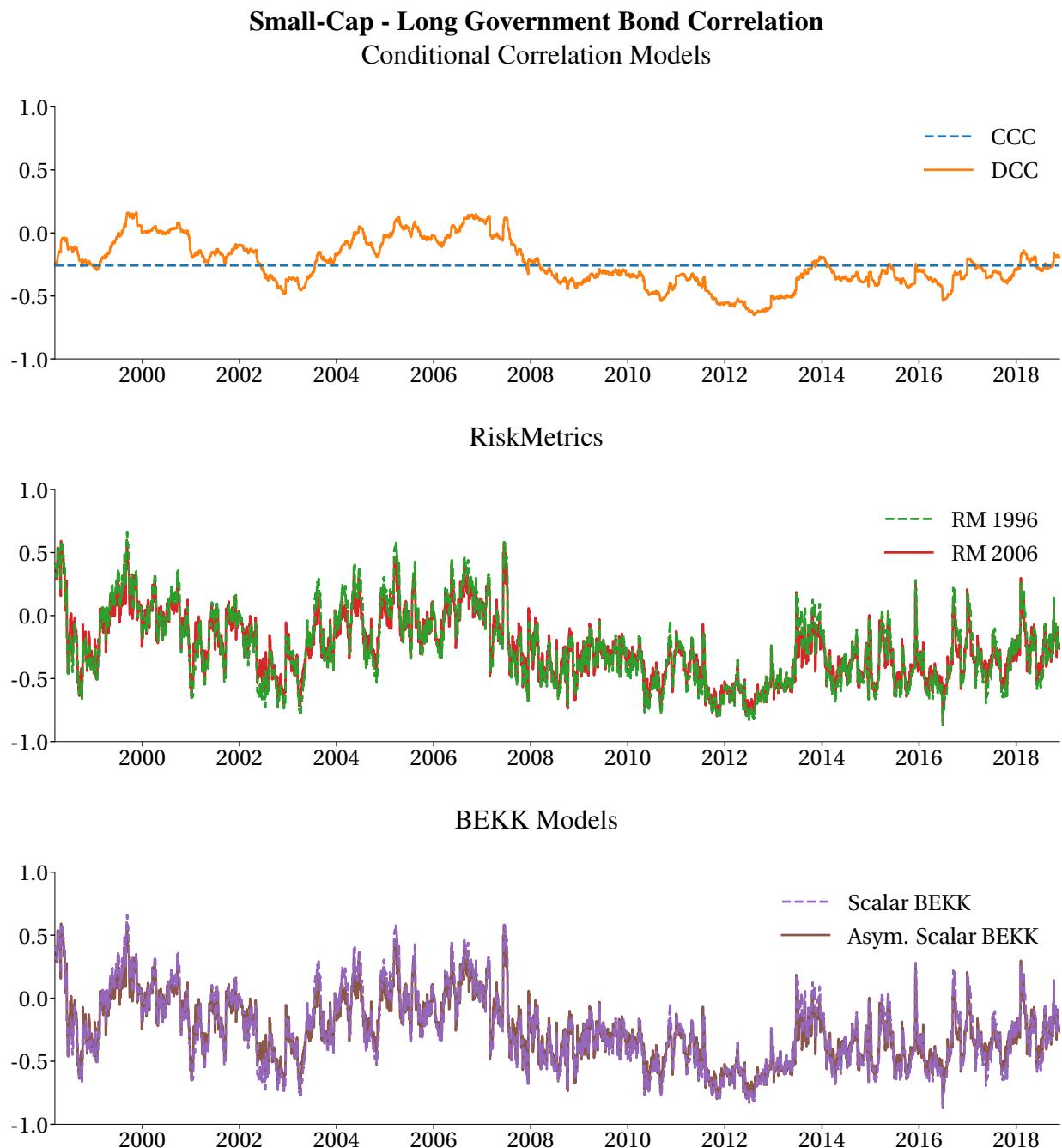


Figure 9.6: The three panels show the estimated conditional correlation between the small-cap fund and the bond fund from 6 models.

Prices should, in principle, be sampled as frequently as possible to maximize the precision of the Realized Covariance estimator. In practice, frequent sampling is not possible since:

- Prices, especially transaction prices (trades), are contaminated by noise (e.g., bid-ask bounce).
- Prices are not perfectly synchronized. For example, if asset i trades at 10:00:00 and the last

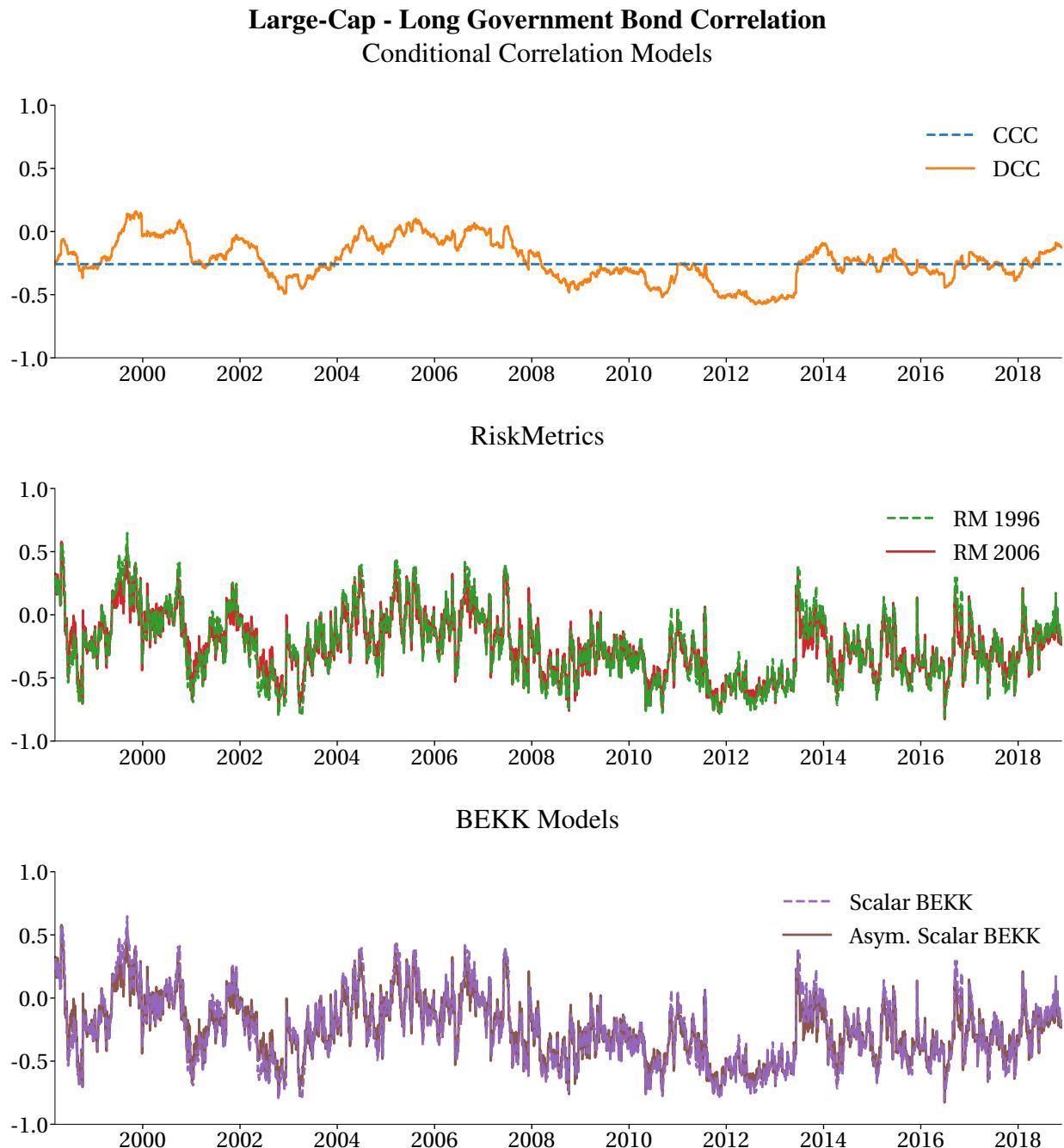


Figure 9.7: The three panels show the estimated conditional correlation between the large-cap fund and the bond fund from 6 models.

trade of asset j occurs at 9:59:50, then the estimated covariance is biased towards 0.

The conventional method to address these two concerns is to sample relatively infrequently, for example, every 5 minutes.

The standard Realized Covariance estimator can be improved using *subsampling*. For example,

suppose prices are available every minute, but that microstructure concerns (noise and synchronization) limit sampling to 10-minute returns. The subsampled Realized Covariance uses *all* 10-minute returns, not just non-overlapping ones, to estimate the covariance.

Definition 9.23 (Subsampled Realized Covariance). The subsampled Realized Covariance estimator is defined

$$\begin{aligned} RC_{t,SS}^{(m,n)} &= \frac{m}{n(m-n+1)} \sum_{i=1}^{m-n+1} \sum_{j=1}^n \mathbf{r}_{i+j-1,t} \mathbf{r}'_{i+j-1,t} \\ &= \frac{1}{n} \sum_{j=1}^n \frac{m}{(m-n+1)} \sum_{i=1}^{m-n+1} \mathbf{r}_{i+j-1,t} \mathbf{r}'_{i+j-1,t} \\ &= \frac{1}{n} \sum_{j=1}^n \widetilde{RC}_{j,t}, \end{aligned} \quad (9.52)$$

where there are m high-frequency returns available and the selected sampling time is based on n returns.

For example, suppose data is available from 9:30:00 to 16:00:00, and that prices are sampled every minute. The standard Realized Covariance uses returns constructed from prices sampled at 9:30:00, 9:40:00, 9:50:00, The subsampled Realized Covariance uses returns computed from all 10-minute windows, i.e., 9:30:00 and 9:40:00, 9:31:00 and 9:41:00, 9:32:00 and 9:42:00, and so on. In this example, m is the number of 1-minute returns available over a 6.5 hour day (390), and n is the number of 1-minute returns in the desired sampling frequency of 10-minutes (10).

Barndorff-Nielsen, Hansen, Lunde, and Shephard (2011) propose an alternative method to estimate the integrated covariance known as a *Realized Kernel*. It is superficially similar to Realized Covariance except that Realized Kernels use a weighting function similar to that in the Newey and West (1987) covariance estimator.

Definition 9.24 (Realized Kernel). The Realized Kernel is defined as

$$\begin{aligned} RK_t &= \Gamma_0 + \sum_{i=1}^h K\left(\frac{i}{H+1}\right) (\Gamma_i + \Gamma'_i) \\ \Gamma_j &= \sum_{i=j+1}^{\tilde{m}} \tilde{\mathbf{r}}_{i,t} \tilde{\mathbf{r}}'_{i-j,t} \end{aligned} \quad (9.53)$$

where $\tilde{\mathbf{r}}$ are refresh time returns, \tilde{m} is the number of refresh time returns, $K(\cdot)$ is a kernel weighting function and H is a parameter which controls the bandwidth.

Refresh time returns are needed to ensure that prices are not overly stale, and are constructed by sampling prices using last-price interpolation only after all assets have traded. For example, Table 9.5 contains a set of simulated trade times for SPY, a leading ETF that tracks the S&P 500, and GLD, an ETF that tracks the price of gold. A tick (✓) indicates that a trade occurs at the timestamp in the first column. A tick in the refresh column indicates that this timestamp is a refresh time. The final two columns contain the timestamps of the prices used to compute the refresh-time returns.

Trade Time	SPY	GLD	Refresh	SPY Time	GLD Time
9:30:00	✓	✓	✓	9:30:00	9:30:00
9:30:01	✓	✓	✓	9:30:01	9:30:01
9:30:02					
9:30:03	✓				
9:30:04	✓				
9:30:05		✓	✓	9:30:04	9:30:05
9:30:06		✓			
9:30:07		✓			
9:30:08	✓		✓	9:30:08	9:30:07

Table 9.5: This table illustrates refresh-time price construction. Prices are sampled after all assets have traded using last-price interpolation. Refresh-time sampling usually eliminated some of the data, e.g., the 9:30:03 trade of SPY and prices are not perfectly synchronized, e.g., the 9:30:08 refresh-time price which uses the SPY price from 9:30:08 and the GLD price from 9:30:07.

The recommended kernel is Parzen's kernel,

$$K(x) = \begin{cases} 1 - 6x^2 + 6x^3 & 0 > x \geq \frac{1}{2} \\ 2(1-x)^3 & \frac{1}{2} > x \geq 1 \\ 0 & x > 1 \end{cases} \quad (9.54)$$

The bandwidth parameter, H , plays a crucial role in the accuracy of Realized Kernels. A discussion of the estimation of the bandwidth is beyond the scope of these notes. See Barndorff-Nielsen, Hansen, Lunde, and Shephard (2008) and Barndorff-Nielsen, Hansen, Lunde, and Shephard (2011) for detailed discussions.

9.5.1 Realized Correlation and Beta

Realized Correlation is the realized analog of the usual correlation estimator, and is defined using the Realized Covariance.

Definition 9.25 (Realized Correlation). The realized correlation between two series is defined

$$RCorr = \frac{RC_{ij}}{\sqrt{RC_{ii}RC_{jj}}}$$

where RC_{ij} is the Realized Covariance between assets i and j and RC_{ii} and RC_{jj} are the realized variances of assets i and j, respectively.

Realized Betas are similarly defined, only using the definition of a regression β (which is a function of the covariance).

Definition 9.26 (Realized Beta). Suppose RC_t is a $k+1$ by $k+1$ Realized Covariance matrix for an asset and a set of observable factors where the asset is in position 1, so that the Realized Covariance

can be partitioned

$$RC = \begin{bmatrix} RC_{ii} & RC'_{fi} \\ RC_{fi} & RC_{ff} \end{bmatrix}$$

where RC_{ii} is the Realized Variance of the asset, RC_{if} is the k by 1 vector of Realized Covariance between the asset and the factors, and RC_{ff} is the k by k Realized Covariance of the factors. The Realized Beta is defined

$$R\beta = RC_{ff}^{-1}RC_{fi}.$$

In the usual case where there is only one factor, usually the market, the realized beta is the ratio of the Realized Covariance between the asset and the market to the variance of the market. Realized Betas are similar to other realized measures in that they are model free and, as long as prices can be sampled frequently and have little market microstructure noise, are accurate measures of the exposure to changes in the market.

9.5.2 Modeling Realized Covariance

Modified multivariate ARCH models can be used to modeling and forecast Realized Covariance and Realized Kernels. The basic assumption is that the *mean* of the Realized Covariance, conditional on the time $t - 1$ information, is Σ_t ,

$$RC_t | \mathcal{F}_{t-1} \sim F(\Sigma_t, v) \quad (9.55)$$

where $F(\cdot, \cdot)$ is some distribution with conditional mean Σ_t which may depend on other parameters unrelated to the mean which are contained in v . This assumption implies that the Realized Covariance is driven by a matrix-valued shock which has conditional expectation \mathbf{I}_k ,

$$RC_t = \Sigma_t^{\frac{1}{2}} \Xi \Sigma_t^{\frac{1}{2}}$$

where $\Xi \stackrel{\text{i.i.d.}}{\sim} F(\mathbf{I}, \tilde{v})$ and \tilde{v} is used to denote that these parameters are related to but different from those in eq. 9.55. This assumption is identical to the one made when modeling realized variance as a non-negative process with a multiplicative error (MEM) where it is assumed that $RV_t = \sigma_t^2 \xi_t = \sigma_t \xi_t \sigma_t$ where $\xi_t \stackrel{\text{i.i.d.}}{\sim} F(1, v)$.

Most multivariate ARCH models can be adapted by replacing the outer product of the shocks with the Realized Covariance. For example, consider the standard BEKK model,

$$\Sigma_t = \mathbf{C}\mathbf{C}' + \mathbf{A}\mathbf{r}_{t-1}\mathbf{r}_{t-1}\mathbf{A}' + \mathbf{B}\Sigma_{t-1}\mathbf{B}'.$$

The BEKK can be viewed as a multiplicative error model and used for Realized Covariance by specifying the dynamics as

$$\Sigma_t = \mathbf{C}\mathbf{C}' + \mathbf{A}RC_{t-1}\mathbf{A}' + \mathbf{B}\Sigma_{t-1}\mathbf{B}'.$$

Other ARCH models can be similarly adapted by replacing the outer product of returns by the Realized Covariance or Realized Kernel. Estimation is no more difficult than the estimation of the parameters in a multivariate ARCH model, and the parameters can be estimated by maximizing the Wishart log-likelihood. See Noureldin, Shephard, and Sheppard (2012) for details.

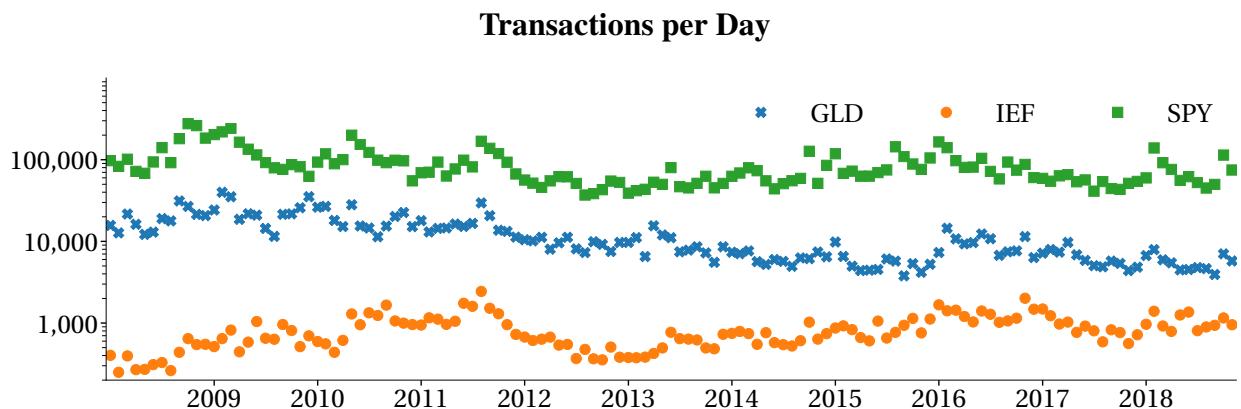


Figure 9.8: The average number of daily transactions in each month of the sample for the three ETFs: SPDR S&P 500 ETF, SPDR Gold Trust (GLD), and iShares 7-10 Year Treasury Bond ETF.

9.5.3 Application: ETF Realized Covariance

Exchange-traded funds have emerged as popular instruments that facilitate investing in assets that are often difficult to access for retail investors. They trade like stocks but are backed by other assets or derivative securities. Three ETFs are used to highlight some of the issues unique to Realized Covariance that are not important when modeling a single asset. The funds used are the SPDR S&P 500 ETF (SPY), which tracks the S&P 500, SPDR Gold Trust (GLD), which aims to track to the spot price of gold, and iShares 7-10 Year Treasury Bond ETF (IEF), which tracks the return on intermediate maturity U.S. government debt. The data used in this application run from the start of 2008 until the end of 2018. The estimators are implemented using only transaction data (trades) that are available during the normal trading hours of 9:30 (Eastern/US) to 16:00.

Figure 9.8 shows the average number of transactions per day for the three ETFs. There are substantial differences in the liquidity of the three funds. IEF trades about 800 times per day, on average, over the sample. In some months, the average number of transaction is as low as 250, while in periods of higher liquidity the fund is traded over 1,000 times per day. The S&P 500-tracking ETF consistently trades over 80,000 times per day. The U.S. trading day last 6.5 hours and so the time between trades ranges between 30 and 90 seconds for IEF and is less than half a second for SPY. The liquidity of the least liquid asset always serves an upper bound for the sampling frequency used when estimating RC . In this application, sampling more frequently than 30 seconds is likely to produce a sharp reduction in covariance and correlation for pairs involving IEF. GLD's liquidity is consistently between IEF and SPY and trades typically occur every 3 seconds.

Figure 9.9 contains two signature plots. The top is known as the pseudo-correlation signature and plots the time-averaged Realized Covariance standardized by the average cross-product of realized volatilities sampled at a single (conservative) frequency.

Definition 9.27 (Pseudo-Correlation Signature Plot). The pseudo-correlation signature plot displays the time-series average of Realized Covariance

$$\overline{RCorr}_{ij,t}^{(m)} = \frac{T^{-1} \sum_{t=1}^T RC_{ij,t}^{(m)}}{\overline{RVol}_i \times \overline{RVol}_j}$$

where m is the number of samples and $\overline{RVol}_\bullet = \sqrt{T^{-1} \sum_{t=1}^T RC_{\bullet\bullet,t}^{(q)}}$ is the square root of the average Realized Variance using q -samples. q is chosen to produce an accurate RV that is free from microstructure effects. An equivalent representation displays the amount of time, either in calendar time or tick time (number of trades between observations) along the x-axis.

The pseudo-correlations all diverge from 0 as the sampling interval grows. The pseudo-correlation between the Gold and the S&P 500 ETFs appears to reach its long-run level when sampling prices every 2 minutes. This sampling interval is surprisingly long considering that the slower of these two assets, GLD, trades about every 3 seconds on average. The pseudo-correlations involving the U.S. bond ETF continue to move away from 0 until the sample interval is 10 minutes, which reflects the lower liquidity in this ETF.

The slow convergence of both series is known as the Epps Effect (Epps, 1979). Epps first documented that correlations converge to 0 as the sampling frequency increases. There are two reasons why the correlations converge to zero as the sampling frequency increases: the numerator (covariance) reducing in magnitude or the denominator increasing due to bid-ask bounce. The bottom panel of Figure 9.9 plots the annualized cross-volatility signature of the two series. The cross-volatility signatures are remarkably flat, and so the changes in the pseudo-correlation signature are due to changes in the covariances.

Definition 9.28 (Cross-volatility Signature Plot). The cross-volatility signature plot displays the square-root of the time-series average of the product of two Realized Variances,

$$\overline{XVol}_{ij,t}^{(m)} = \sqrt{T^{-1} \sum_{t=1}^T RV_{i,t}^{(m)} \times T^{-1} \sum_{t=1}^T RV_{j,t}^{(m)}}$$

where m is the number of samples and $RV_{\bullet,t}^{(m)} = RC_{\bullet\bullet,t}^{(m)}$ are the diagonal elements of the Realized Covariance matrix. An equivalent representation displays the amount of time, whether in calendar time or tick time (number of trades between observations) along the X-axis. It is often presented in annualized terms,

$$\text{Ann. } \overline{XVol}_{ij,t}^{(m)} = \sqrt{252 \times \overline{XVol}_{ij,t}^{(m)}}.$$

The pseudo-correlation signature plot can be misleading if covariance does not consistently have the same sign. For example, suppose two assets have a long-run correlation near 0 but have persistent deviations where their correlation is either positive or negative for long periods. The pseudo-correlation signature may appear flat for all sampling times even though the correlation is not well estimated. An alternative is to use a R^2 -signature plot which is defined by transforming the Realized Covariances into the Realized β and an idiosyncratic variance. In the model $Y_i = \alpha + \beta X_i + \varepsilon_i$, the variance of the idiosyncratic residual is $V[Y] - \beta^2 V[X]$. Scaling this variance by the variance of Y produces $\frac{V[Y] - \beta^2 V[X]}{V[Y]} = 1 - R^2$.

The R^2 signature plot displays the scaled average residual variance,

$$\overline{R^2}_{ij,t}^{(m)} = 1 - \frac{T^{-1} \sum_{t=1}^T RC_{ii,t}^{(m)} - (RC_{ij,t}^{(m)})^2 / RC_{jj,t}^{(m)}}{T^{-1} \sum_{t=1}^T RC_{ii,t}^{(m)}} = 1 - \frac{T^{-1} \sum_{t=1}^T RC_{ii,t}^{(m)} - (R\beta_{ij,t}^{(m)})^2 / RC_{jj,t}^{(m)}}{T^{-1} \sum_{t=1}^T RC_{ii,t}^{(m)}}$$

where $RC_t^{(m)}$ is the m -sample Realized Covariance in month j .

The R^2 could be low if the variance components were large, which may happen if the returns are contaminated by market microstructure noise, or if the β is not accurately measured.

Finally, Figure 9.10 shows the estimated correlations of these ETFs. The dashed line shows the average correlation computed by transforming the time-averaged Realized Covariance to a correlation. The S&P 500 tracking ETF is negativity correlated, on average, with both Gold and U.S. Treasuries. The correlation is unusually low near the start of 2016 and is also near the bottom of its range at the end of 2018. Gold and Treasuries, on the other hand, were highly correlated in 2017 and only returned to their long-run level towards the end of the sample.

9.6 Measuring Dependence

Covariance does not completely characterize the dependence between asset returns. It only measures the linear dependence between the returns and so may be misleading if assets have nonlinear relationships.

9.6.1 Linear Dependence

Linear or Pearson correlation is the most common measure of dependence.

Definition 9.29 (Linear (Pearson) Correlation). The linear (Pearson) correlation between two random variables X and Y is

$$\rho = \frac{\text{Cov}[X, Y]}{\sqrt{\text{V}[X] \text{V}[Y]}}. \quad (9.56)$$

The sample correlation estimator is

$$\hat{\rho} = \frac{\sum_{t=1}^T (X_t - \hat{\mu}_x)(Y_t - \hat{\mu}_y)}{\sqrt{\sum_{t=1}^T (Y_t - \hat{\mu}_x)^2 \sum_{s=1}^T (Y_s - \hat{\mu}_y)^2}}. \quad (9.57)$$

where $\hat{\mu}_x$ and $\hat{\mu}_y$ are the sample means of X_t and Y_t , respectively.

Linear correlation measures the strength of the linear relationship between standardized versions of X and Y . Correlation is invariant to affine increasing transformations of X or Y (i.e., $a + bY, b > 0$). It is not, however, invariant to non-linear transformations, even when the non-linear transformation is order preserving (e.g., the log of a non-negative random variable). Linear correlation is also insufficient to characterize the dependence between two random variables, except when X and Y follow a bivariate normal distribution. Moreover, two distributions can have the same correlation yet have different behavior during extreme events.

9.6.2 Non-linear Dependence

Many measures have been designed to overcome the shortcomings of linear correlation as a measure of risk. These are broadly classified as measures of non-linear dependence.

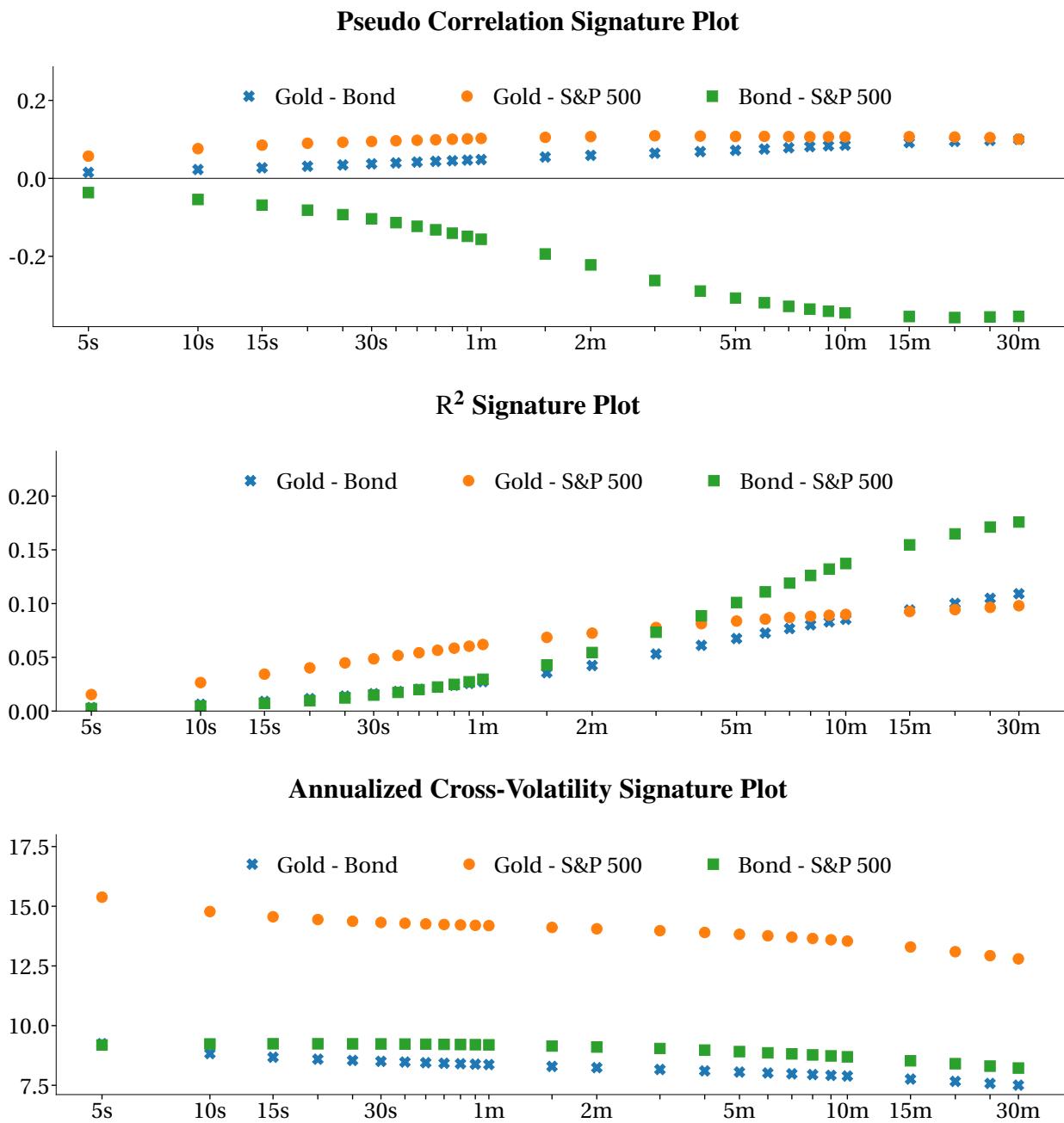


Figure 9.9: The top panel contains the pseudo-correlation signature for the three assets pairs defined as the ratio of the average covariance sampled at different frequencies standardized by a single, fixed-sampling interval cross-product of volatilities. The middle panel plots an alternative sign-free signature plot constructed by squaring the realized correlations. The bottom plot shows the average annualized cross-volatility for average Realized Variances sampled at different frequencies.

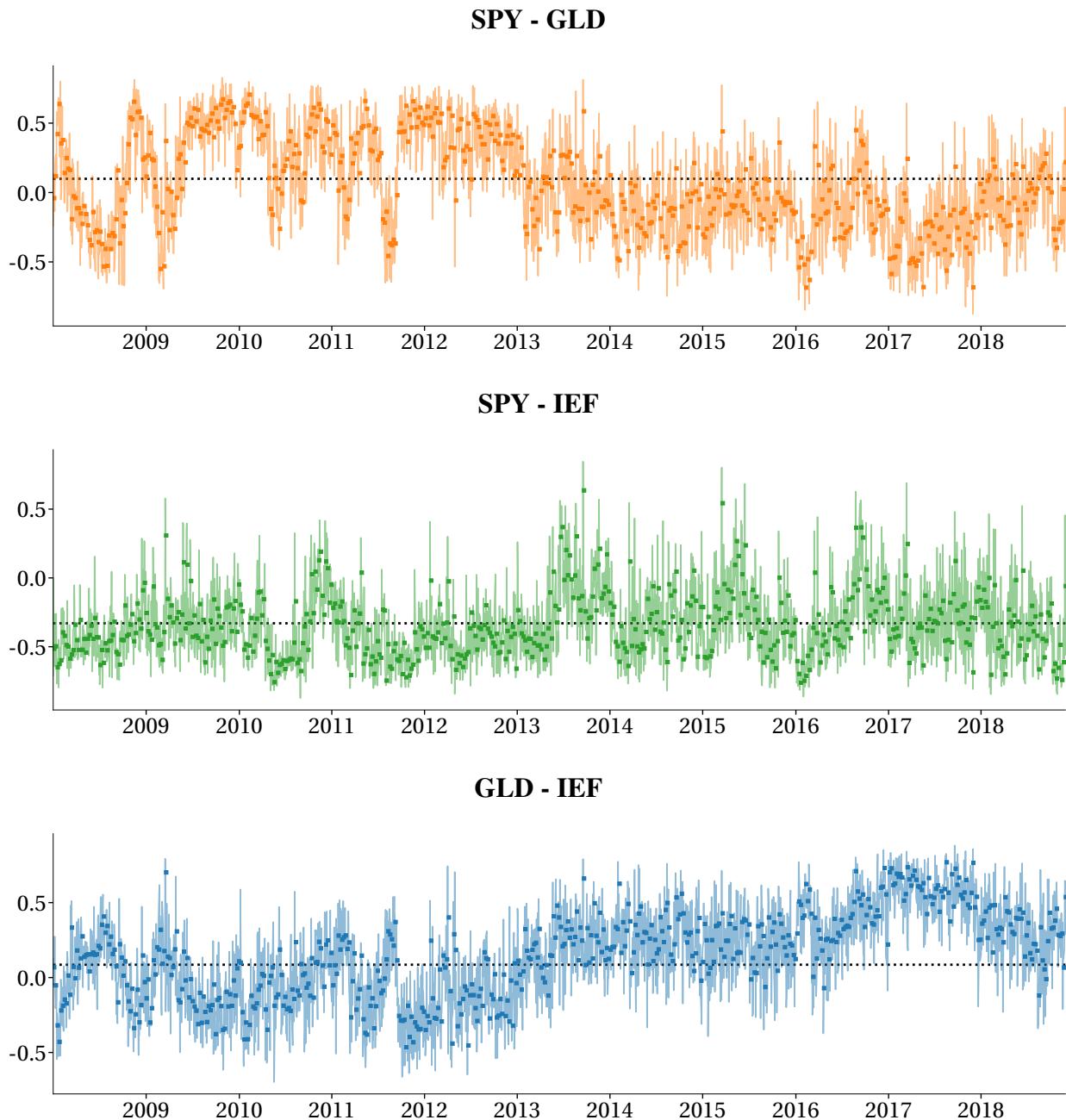


Figure 9.10: Plot of the Realized Correlations between the three ETFs: SPDR S&P 500, SPDR Gold Trust, and iShares 7-10 Year Treasury Bond ETF. All realized correlations are estimated using RC^{SS} based on 15-minute returns subsampled from prices sampled every 5 seconds ($m = 4,680$, $n = 26$). The markers show the weekly realized correlation computed from weekly-averaged Realized Covariances.

9.6.2.1 Rank Correlation

Rank correlation, also known as Spearman correlation, is an alternative measure of dependence which can assess the strength of a relationship and is robust to certain non-linear transformations. Suppose X and Y are random variables, $X \stackrel{\text{i.i.d.}}{\sim} N(0, 1)$ and $Y \equiv X^\lambda$ where λ is odd. If $\lambda = 1$ then $Y = X$ and the linear correlation is 1. If $\lambda = 3$ the correlation is .77. If $\lambda = 5$ then the correlation is only .48, despite Y being a function of only X . As λ increases, the correlation becomes arbitrarily small despite the perfect dependence between X and Y . Rank correlation is robust to increasing non-linear transformations, and the rank correlation between X and Y is 1 for any odd power λ .

Definition 9.30 (Rank (Spearman) Correlation). The rank (Spearman) correlation between two random variables X and Y is

$$\rho_s(X, Y) = \text{Corr}(F_X(X), F_Y(Y)) = \frac{\text{Cov}[F_X(X), F_Y(Y)]}{\sqrt{\text{V}[F_X(X)] \text{V}[F_Y(Y)]}} = 12\text{Cov}[F_X(X), F_Y(Y)] \quad (9.58)$$

where the final identity uses the fact that the variance of a Uniform(0, 1) is $\frac{1}{12}$.

The rank correlation measures the correlation between the *probability integral transforms* of X and Y . The use of the probability integral transform means that rank correlation is preserved under strictly increasing transformations (decreasing monotonic changes the sign), and so $\rho_s(X, Y) = \rho_s(T_1(X), T_2(Y))$ when T_1 and T_2 are any strictly increasing functions.

The sample analog of the Spearman correlation makes use of the empirical ranks of the observed data. Define $R_{X,i}$ to be the rank of X_i , where a rank of 1 corresponds to the smallest value, a rank of n corresponds to the largest value, where any ties are all assigned the average value of the ranks associated with the values in the tied group. Define $R_{Y,i}$ in an identical fashion on Y_i . The sample rank correlation between X and Y is computed as the sample correlation of the ranks,

$$\hat{\rho}_s = \frac{\sum_{i=1}^n \left(\frac{R_{X,i}}{n+1} - \frac{1}{2} \right) \left(\frac{R_{Y,i}}{n+1} - \frac{1}{2} \right)}{\sqrt{\sum_{i=1}^n \left(\frac{R_{X,i}}{n+1} - \frac{1}{2} \right)^2} \sqrt{\sum_{j=1}^n \left(\frac{R_{Y,j}}{n+1} - \frac{1}{2} \right)^2}} = \frac{\sum_{i=1}^n \left(R_{X,i} - \frac{n+1}{2} \right) \left(R_{Y,i} - \frac{n+1}{2} \right)}{\sqrt{\sum_{i=1}^n \left(R_{X,i} - \frac{n+1}{2} \right)^2} \sqrt{\sum_{j=1}^n \left(R_{Y,j} - \frac{n+1}{2} \right)^2}}$$

where $\frac{R_{X,i}}{n+1}$ is the empirical quantile of X_i .

9.6.2.2 Kendall's τ

Kendall's τ is an alternative measure of non-linear dependence which is based on the idea of concordance. Concordance is defined using the signs of pairs of random variables.

Definition 9.31 (Concordant Pair). The pairs of random variables (X_i, Y_i) and (X_j, Y_j) are concordant if $\text{sgn}(X_i - X_j) = \text{sgn}(Y_i - Y_j)$ where $\text{sgn}(\cdot)$ is the sign function which returns -1 for negative values, 0 for zero, and +1 for positive values (equivalently defined as $\text{sgn}((X_i - X_j)(Y_i - Y_j))$).

If a pair is not concordant, then it is *discordant*.

Definition 9.32 (Kendall's τ). Kendall τ is defined

$$\tau = \Pr(\text{sgn}(X_i - X_j) = \text{sgn}(Y_i - Y_j)) - \Pr(\text{sgn}(X_i - X_j) \neq \text{sgn}(Y_i - Y_j)) \quad (9.59)$$

Dependence Measures for Weekly FTSE and S&P 500 Returns

Linear (Pearson)	0.678 (0.027)	Rank (Spearman)	0.613 (0.031)	Kendall's τ	0.446 (0.027)
------------------	------------------	-----------------	------------------	------------------	------------------

Table 9.6: Linear and rank correlation and Kendall's τ (bootstrap std. error in parenthesis) for weekly returns for the S&P 500 and FTSE 100.

The estimator of Kendall's τ uses the sample analogs to the probabilities in the definition. Defined $n_c = \sum_{i=1}^n \sum_{j=i+1}^n I_{[\text{sgn}(X_i - X_j) = \text{sgn}(Y_i - Y_j)]}$ as the count of the concordant pairs and $n_d = \frac{1}{2}n(n-1) - n_c$ as the count of discordant pairs. The estimator of τ is

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \quad (9.60)$$

$$= \frac{n_c}{\frac{1}{2}n(n-1)} - \frac{n_d}{\frac{1}{2}n(n-1)} \quad (9.61)$$

$$= \widehat{\Pr}(\text{sgn}(X_i - X_j) = \text{sgn}(Y_i - Y_j)) - \widehat{\Pr}(\text{sgn}(X_i - X_j) \neq \text{sgn}(Y_i - Y_j)) \quad (9.62)$$

where $\widehat{\Pr}$ denotes the empirical probability. Kendall's τ measures the difference between the probability a pair is concordant, $n_c/(\frac{1}{2}n(n-1))$ and the probability a pair is discordant $n_d/(\frac{1}{2}n(n-1))$. Since τ is the difference between two probabilities it must fall in $[-1, 1]$ where -1 indicates that all pairs are discordant, 1 indicates that all pairs are concordant, and τ is increasing as the concordance between the pairs increases. Like rank correlation, Kendall's τ is also invariant to increasing transformation since a pair that is concordant before the transformation (i.e., $X_i > X_j$ and $Y_i > Y_j$) is also concordant after a strictly increasing transformation (i.e., $T_1(X_i) > T_1(X_j)$ and $T_2(Y_i) > T_2(Y_j)$).

9.6.2.3 Exceedance Correlations and Betas

Exceedance correlation, like expected shortfall, is one of many exceedance measures which can be constructed by computing expected values conditional on exceeding some threshold. Exceedance correlation measures the correlation between the variables *conditional* on both variables taking values in their upper or lower tail.

Definition 9.33 (Exceedance Correlation). The exceedance correlation at level κ is defined as

$$\rho^+(\kappa) = \text{Corr}[X, Y | X > \kappa, Y > \kappa] \quad (9.63)$$

$$\rho^-(\kappa) = \text{Corr}[X, Y | X < -\kappa, Y < -\kappa] \quad (9.64)$$

Exceedance correlation is computed using the standard (linear) correlation estimator on the subset of data where both $X > \kappa$ and $Y > \kappa$ (positive) or $X < -\kappa$ and $Y < -\kappa$. Exceedance correlation can also be defined using series specific cutoff points such as κ_X and κ_Y , which are often used if the series do not have the same variance. Series-specific thresholds are often set using quantiles of X and Y (e.g., the 10% quantile of each). Alternatively, exceedance correlations can be computed with data transformed to have unit variance. Sample exceedance correlations are computed as

$$\hat{\rho}^+(\kappa) = \frac{\hat{\sigma}_{xy}^+(\kappa)}{\hat{\sigma}_x^+(\kappa)\hat{\sigma}_y^+(\kappa)}, \quad \hat{\rho}^-(\kappa) = \frac{\hat{\sigma}_{xy}^-(\kappa)}{\hat{\sigma}_x^-(\kappa)\hat{\sigma}_y^-(\kappa)} \quad (9.65)$$

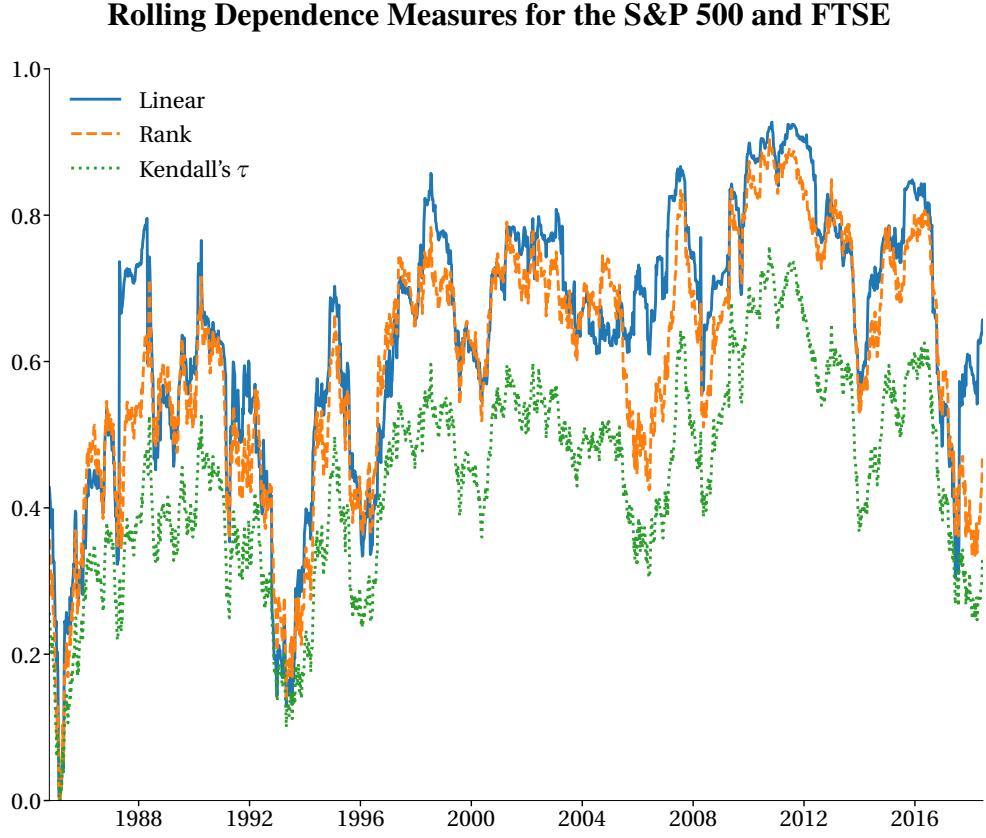


Figure 9.11: Plot of rolling linear correlation, rank correlation and Kendall's τ between weekly returns on the S&P 500 and the FTSE estimated using 1-year moving windows. The measures broadly agree about the changes in dependence but not the level.

where

$$\begin{aligned}
 \hat{\sigma}_{XY}^+(\kappa) &= \frac{\sum_{t=1}^T (X_t - \mu_X^+(\kappa)) (Y_t - \mu_Y^+(\kappa)) I_{[X_t > \kappa \cap Y_t > \kappa]}}{T_\kappa^+} \\
 \hat{\sigma}_{XY}^-(\kappa) &= \frac{\sum_{t=1}^T (X_t - \mu_X^-(\kappa)) (Y_t - \mu_Y^-(\kappa)) I_{[X_t < -\kappa \cap Y_t < -\kappa]}}{T_\kappa^-} \\
 \hat{\mu}_X^+(\kappa) &= \frac{\sum_{t=1}^T X_t I_{[X_t > \kappa \cap Y_t > \kappa]}}{T_\kappa^+}, \quad \hat{\sigma}_X^{2+}(\kappa) = \frac{\sum_{t=1}^T (X_t - \hat{\mu}_X^+(\kappa))^2 I_{[X_t > \kappa \cap Y_t > \kappa]}}{T_\kappa^+} \\
 \hat{\mu}_X^-(\kappa) &= \frac{\sum_{t=1}^T X_t I_{[X_t < -\kappa \cap Y_t < -\kappa]}}{T_\kappa^-}, \quad \hat{\sigma}_X^{2-}(\kappa) = \frac{\sum_{t=1}^T (X_t - \hat{\mu}_X^-(\kappa))^2 I_{[X_t < -\kappa \cap Y_t < -\kappa]}}{T_\kappa^-} \\
 T_\kappa^+ &= \sum_{t=1}^T I_{[X_t > \kappa \cap Y_t > \kappa]}, \quad T_\kappa^- = \sum_{t=1}^T I_{[X_t < -\kappa \cap Y_t < -\kappa]}
 \end{aligned}$$

where the quantities for Y are similarly defined. Exceedance correlation can only be estimated if the region where $X < \kappa$ and $Y < \kappa$ is populated with data, and it is possible for some assets that this region

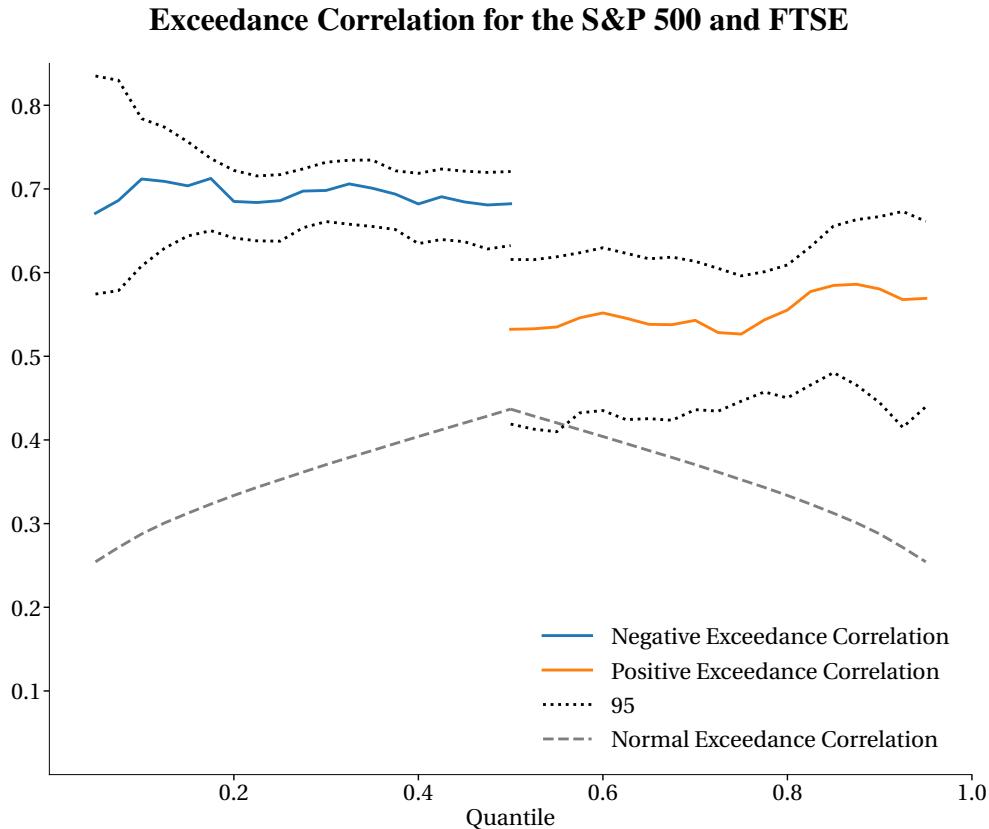


Figure 9.12: Plot of the exceedance correlations with 95% bootstrap confidence intervals for weekly returns on the S&P 500 and FTSE (each series is divided by its sample standard deviation). There is a substantial asymmetry between the positive and negative exceedance correlations.

is empty. Empty regions may occur when measuring the exceedance correlation between assets that have strong negative dependence (e.g., equity and bond returns).

Inference can be conducted using the bootstrap or using analytical methods. Hong, Tu, and Zhou (2007) show that inference on exceedance correlations can be conducted by viewing these estimators as method of moments estimators. Define the standardized exceedance residuals as,

$$\begin{aligned}\tilde{X}_t^+(\kappa) &= \frac{X_t - \mu_x^+(\kappa)}{\sigma_x^+(\kappa)}, & \tilde{X}_t^-(\kappa) &= \frac{X_t - \mu_x^-(\kappa)}{\sigma_x^-(\kappa)}, \\ \tilde{Y}_t^+(\kappa) &= \frac{Y_t - \mu_y^+(\kappa)}{\sigma_y^+(\kappa)}, & \tilde{Y}_t^-(\kappa) &= \frac{Y_t - \mu_y^-(\kappa)}{\sigma_y^-(\kappa)}.\end{aligned}$$

These form the basis of the moment conditions,

$$\begin{aligned}\frac{T}{T_\kappa^+} (\tilde{X}_t^+(\kappa) \tilde{Y}_t^+(\kappa) - \rho^+(\kappa)) I_{[X_t > \kappa \cap Y_t > \kappa]} \\ \frac{T}{T_\kappa^-} (\tilde{X}_t^-(\kappa) \tilde{Y}_t^-(\kappa) - \rho^-(\kappa)) I_{[X_t < -\kappa \cap Y_t < -\kappa]}. \end{aligned}\tag{9.66}$$

Inference on a vector of exceedance correlation can be conducted by stacking the moment conditions and using a HAC covariance estimator such as the Newey and West (1987) estimator. Suppose κ is a vector of thresholds $\kappa_1, \kappa_2, \dots, \kappa_n$, then

$$\sqrt{T} \begin{pmatrix} \hat{\rho}^+(\kappa) - \rho^+(\kappa) \\ \hat{\rho}^-(\kappa) - \rho^-(\kappa) \end{pmatrix} \xrightarrow{d} N(\mathbf{0}, \Omega)$$

Ω can be estimated using the moment conditions,

$$\hat{\Omega} = \hat{\Gamma}_0 + \sum_{l=1}^L w_l (\hat{\Gamma}_l + \hat{\Gamma}'_l) \quad (9.67)$$

where $w_l = 1 - \frac{l}{L+1}$,

$$\hat{\Gamma}_j = \sum_{t=j+1}^T \xi_t \xi_{t-j}$$

and

$$\xi_t = T \begin{bmatrix} \frac{1}{T_{\kappa_1}^+} (\tilde{X}^+(\kappa_1) \tilde{Y}^+(\kappa_1) - \rho^+(\kappa)) I_{[X_t > \kappa_1 \cap Y_t > \kappa_1]} \\ \vdots \\ \frac{1}{T_{\kappa_n}^+} (\tilde{X}^+(\kappa_n) \tilde{Y}^+(\kappa_n) - \rho^+(\kappa_n)) I_{[X_t > \kappa_n \cap Y_t > \kappa_n]} \\ \frac{1}{T_{\kappa_1}^-} (\tilde{X}^-(\kappa_1) \tilde{Y}^-(\kappa_1) - \rho^-(\kappa)) I_{[X_t > \kappa_1 \cap Y_t > \kappa_1]} \\ \vdots \\ \frac{1}{T_{\kappa_n}^-} (\tilde{X}^-(\kappa_n) \tilde{Y}^-(\kappa_n) - \rho^-(\kappa_n)) I_{[X_t > \kappa_n \cap Y_t > \kappa_n]} \end{bmatrix}.$$

Exceedance beta is similarly defined, only using the ratio of an exceedance covariance to an exceedance variance.

Definition 9.34 (Exceedance Beta). The exceedance beta at level κ is defined as

$$\begin{aligned} \beta^+(\kappa) &= \frac{\text{Cov}(X, Y | X > \kappa, Y > \kappa)}{\text{V}(X | X > \kappa, Y > \kappa)} = \frac{\sigma_Y^+(\kappa)}{\sigma_X^+(\kappa)} \rho^+(\kappa) \\ \beta^-(\kappa) &= \frac{\text{Cov}(X, Y | X < -\kappa, Y < -\kappa)}{\text{V}(X | X < -\kappa, Y < -\kappa)} = \frac{\sigma_Y^-(\kappa)}{\sigma_X^-(\kappa)} \rho^-(\kappa) \end{aligned} \quad (9.68)$$

Sample exceedance betas are computed using the sample analogs,

$$\hat{\beta}^+(\kappa) = \frac{\hat{\sigma}_{XY}^+(\kappa)}{\hat{\sigma}_X^{2+}(\kappa)}, \text{ and } \hat{\beta}^-(\kappa) = \frac{\hat{\sigma}_{XY}^-(\kappa)}{\hat{\sigma}_X^{2-}(\kappa)}, \quad (9.69)$$

and inference can be conducted in an analogous manner to exceedance correlations using a HAC estimator and the moment conditions

$$\begin{aligned} & \frac{T}{T_{\kappa}^+} \left(\frac{\sigma_Y^+(\kappa)}{\sigma_X^+(\kappa)} \tilde{X}^+(\kappa) \tilde{Y}^+(\kappa) - \beta^+(\kappa) \right) I_{[X_t > \kappa \cap Y_t > \kappa]} \\ & \frac{T}{T_{\kappa}^-} \left(\frac{\sigma_Y^-(\kappa)}{\sigma_X^-(\kappa)} \tilde{X}^-(\kappa) \tilde{Y}^-(\kappa) - \beta^-(\kappa) \right) I_{[X_t < -\kappa \cap Y_t < -\kappa]}. \end{aligned} \quad (9.70)$$

9.6.3 Application: Dependence between the S&P 500 and the FTSE 100

Daily data for the entire history of both the S&P 500 and the FTSE 100 is provided by Yahoo! Finance. Table 9.6 contains the three correlations and standard errors computed using the bootstrap where weekly returns are used to bias due to nonsynchronous returns (all overlapping 5-day returns are used to estimate all estimators). The linear correlation is the largest, followed by the rank and Kendall's τ . Figure 9.11 plots these same three measures only using 252-day moving averages. The three measures broadly agree about changes in the level of dependence.

Figure 9.12 plots the negative and positive exceedance correlation along with 95% confidence intervals computed using the bootstrap. The exceedance thresholds are chosen using quantiles of each series. The negative exceedance correlation is computed for thresholds less than or equal to 50%, and positive is computed for thresholds greater than or equal to 50%. The correlation between these markets differs substantially depending on the sign of the returns.

9.6.4 Application: Asymmetric Dependence from Simple Models

Asymmetric dependence can be generated from simple models. The simulated data in both panels of figure 9.13 is from a standard CAP-M calibrated to match a typical S&P 500 stock. The market return is simulated from a standardized t_6 with the same variance as the S&P 500 in the past ten years. The idiosyncratic variance is similarly calibrated to the cross-section of idiosyncratic variances.

The simulated data in the top panel is computed from

$$r_{i,t} = r_{m,t} + \varepsilon_{i,t}$$

where $\varepsilon_{i,t}$ is i.i.d. normally distributed and has the same variance as the average idiosyncratic variance in the cross-section of S&P 500 constituents. The simulated data shown in the bottom panel is generated according to

$$r_{i,t} = r_{m,t} + z_{i,t} \varepsilon_{i,t}$$

where $z_{i,t} = \exp(-10r_{m,t} I_{[r_{m,t} < 0]})$ introduce heteroskedasticity so that the idiosyncratic variance is smaller on days where the market is down. This simple change introduces asymmetric dependence between positive and negative returns.

9.7 Copulas

Copulas are a relatively new tool in financial econometrics that have applications in risk management and credit and derivative pricing. Copulas allow a distribution to be decomposed where the dependence between assets is separated from the marginal distribution of each asset. Recall that a k -variate

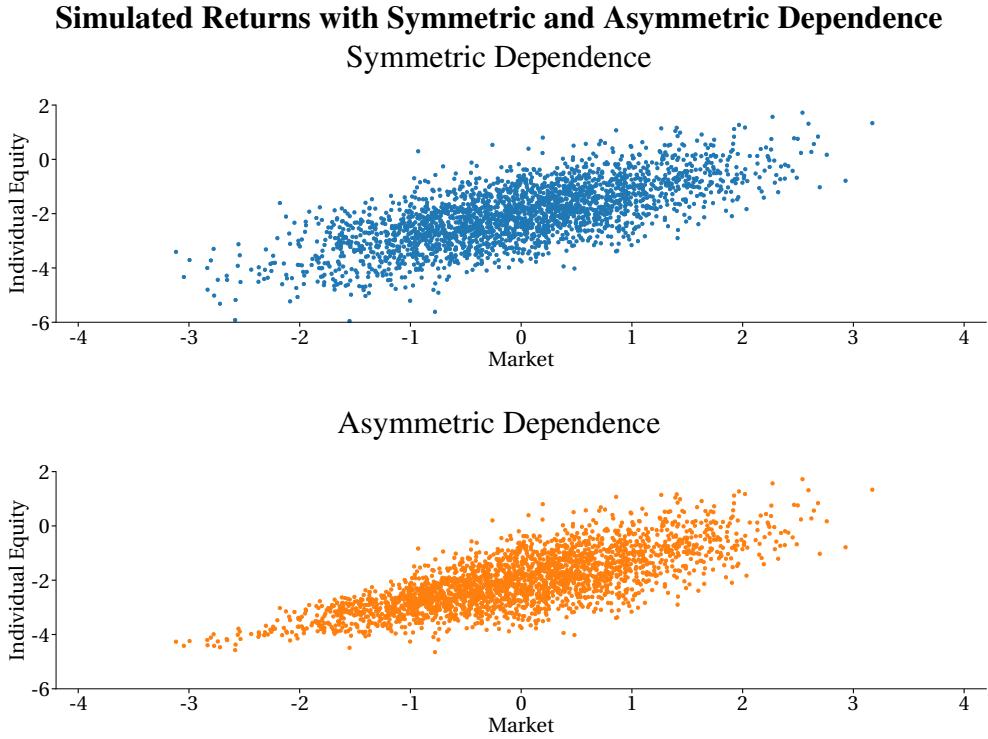


Figure 9.13: These graphs show simulated returns from a CAP-M where the market has a t_6 distribution with the same variance as the S&P 500. The idiosyncratic shock is normally distributed with mean 0, and its variance matches the variance of the idiosyncratic errors of the S&P 500 constituents. The asymmetric dependence is introduced through idiosyncratic error heteroskedasticity where the error variance is $\sigma_\epsilon \exp\left(\frac{1}{2}r_m I_{[r_m < 0]}\right)$. The idiosyncratic component has a smaller variance when the market return is negative than when the market return is positive.

random variable \mathbf{X} has a cumulative distribution function $F(x_1, x_2, \dots, x_k)$ which maps from the domain of \mathbf{X} to $[0, 1]$. The distribution function contains all of the information about the probability of observing different values of \mathbf{X} , and while there are many distribution functions, most are fairly symmetric and rigid. For example, the multivariate Student's t requires all margins to have the same degree-of-freedom parameter, and so the chance of seeing extreme returns – more than 3σ away from the mean – must be the same for all assets. While this assumption may be reasonable when modeling equity index returns, extremely heavy tails are not plausible in other asset classes, e.g., bond or foreign exchange returns. Copulas provide a flexible mechanism to model the marginal distributions *separately* from the dependence structure, and so provide a richer framework for specifying multivariate distributions than the standard set of multivariate distribution functions.

Recall the definition of the marginal distribution of X_1 .

Definition 9.35 (Marginal Density). Let $X = (X_1, X_2, \dots, X_k)$ be a k -variate random variable with joint density $f_X(X)$. The marginal density of X_i is defined

$$f_i(x_i) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_X(x) dx_1 \dots dx_{i-1} dx_{i+1} \dots dx_k.$$

The marginal density contains only information about the probability of observing values of X_i . For example, if X is a bivariate random variable with continuous support, then the marginal density of X_1 is

$$f_1(x_1) = \int_{-\infty}^{\infty} f_X(x_1, x_2) dx_2.$$

The marginal distribution,

$$F_1(x_1) = \int_{-\infty}^{x_1} f_1(s) ds,$$

contains all of the information about the probability of observing values of X_1 , and importantly $F_{X_1}(X_1) \sim U(0, 1)$. The transformation removes the information contained in the marginal distribution about the probability of observing different values of X_1 .

This probability integral transformation applies to both X_1 and X_2 , and so $U_1 = F_{X_1}(x_1)$ and $U_2 = F_{X_2}(x_2)$ only information about the dependence between the two random variables. The distribution that describes the dependence is known as a copula, and so applications built with copulas allow information in marginal distributions to be cleanly separated from the dependence between random variables. This decomposition provides a flexible framework for constructing precise models of *both* the marginal distributions and the dependence.

9.7.1 Basic Theory

A copula is a distribution function for a random variable where each margin is uniform $[0, 1]$.

Definition 9.36 (Copula). A k -dimensional copula is a distribution function on $[0, 1]^k$ with standard uniform marginal distributions, and is denoted $C(u_1, u_2, \dots, u_k)$.

All copulas all satisfy four fundamental properties:

- $C(u_1, u_2, \dots, u_k)$ is increasing in each component u_i ;
- $C(0, \dots, u_j, \dots, 0) = 0$;
- $C(1, \dots, u_j, \dots, 1) = u_j$; and
- for all $\mathbf{u} \leq \mathbf{v}$ where inequality holds on a point-by-point basis, the probability of the hypercube bound with corners \mathbf{u} and \mathbf{v} is non-negative.

Sklar's theorem provides the critical insight that explains how a joint distribution is related to its marginal distributions and the copula that link them. (Sklar, 1959).

Theorem 9.2 (Sklar's Theorem). *Let F be a k -variate joint distribution with marginal distributions F_1, F_2, \dots, F_k . Then there exists a copula $C : [0, 1]^k \rightarrow [0, 1]$ such that for all x_1, x_2, \dots, x_k ,*

$$\begin{aligned} F(x_1, x_2, \dots, x_k) &= C(F_1(x_1), F_2(x_2), \dots, F_k(x_k)) \\ &= C(u_1, u_2, \dots, u_k). \end{aligned}$$

Additionally, if the margins are continuous then C is unique.

Sklar's has two important implications. First, it ensures that the copula is unique whenever the margins are continuous, which is usually the case in financial applications. Second, it shows that a copula can be constructed from any distribution function that has known marginal distributions. Suppose $F(x_1, x_2, \dots, x_k)$ is a known distribution function, and that the marginal distribution function of the i^{th} variable is denoted $F_i(\cdot)$. Further assume that the marginal distribution function is invertible, and denote the inverse as $F_i^{-1}(\cdot)$. The copula implicitly defined by F is

$$C(u_1, u_2, \dots, u_k) = F(F_1^{-1}(u_1), F_2^{-1}(u_2), \dots, F_k^{-1}(u_k)).$$

This relationship allows for many standard distribution functions to be used as the basis for a copula, and appears in the definition of the Gaussian and the Student's t copulas.

Copulas are distribution functions for k -variate uniforms, and like all distribution functions they may (or may not) have an associated density. A copula is a k -variate distribution, and so when the copula density exists it can be derived by differentiating the distribution with respect to each component random variable,

$$c(u_1, u_2, \dots, u_k) = \frac{\partial^k C(u_1, u_2, \dots, u_k)}{\partial u_1 \partial u_2 \dots \partial u_k}. \quad (9.71)$$

9.7.2 Tail Dependence

One final measure of dependence, tail dependence, is useful in understanding risks in portfolios and for comparing copulas. Tail dependence is more of a theoretical construction than a measure that is directly estimated (although it is possible to estimate tail dependence).

Definition 9.37 (Tail Dependence). The upper and lower tail dependence, τ^U and τ^L respectively, are defined as the conditional probability of an extreme event,

$$\tau^U = \lim_{u \rightarrow 1^-} \Pr[X > F_X^{-1}(u) | Y > F_Y^{-1}(u)] \quad (9.72)$$

$$\tau^L = \lim_{u \rightarrow 0^+} \Pr[X < F_X(u) | Y < F_Y(u)] \quad (9.73)$$

where the limits are taken from above for τ^U and below for τ^L .

Tail dependence measures the probability X takes an extreme value given Y takes an extreme value. The dependence between a portfolio and assets used as hedges is particularly important when the portfolio suffers a loss day, and so has a return in its lower tail.

Lower tail dependence takes a particularly simple form when working in copulas, and is defined

$$\tau^L = \lim_{u \rightarrow 0^+} \frac{C(u, u)}{u} \quad (9.74)$$

$$\tau^U = \lim_{u \rightarrow 1^-} \frac{1 - 2u + C(u, u)}{1 - u} \quad (9.75)$$

The coefficient of tail dependence is always in $[0, 1]$ since it is a probability. When τ^U (τ^L) is 0, then the two series are upper (lower) tail-independent. When the value is nonzero, the random variables are tail-dependent and higher values indicate more dependence during extreme events.

9.7.3 Copulas

A large number of copulas have been developed. Some, such as the Gaussian, are implicitly defined from standard distributions. Others have been designed only for uniform random variables. In all expressions for the copulas, $U_i \sim U(0, 1)$ are uniform random variables.

9.7.3.1 Independence Copula

The simplest copula is the independence copula which depends only on the product of the input values.

Definition 9.38 (Independence Copula). The independence copula is

$$C(u_1, u_2, \dots, u_k) = \prod_{i=1}^k u_i \quad (9.76)$$

The independence copula has no parameters.

9.7.3.2 Comonotonicity Copula

The copula with the most dependence is known as the comonotonicity copula.

Definition 9.39 (Comonotonicity Copula). The comonotonicity copula is

$$C(u_1, u_2, \dots, u_k) = \min(u_1, u_2, \dots, u_k) \quad (9.77)$$

The dependence in this copula is *perfect*. The comonotonicity does not have an associated copula density.

9.7.3.3 Gaussian Copula

The Gaussian (normal) copula is implicitly defined using the k -variate Gaussian distribution, $\Phi_k(\cdot)$, and the univariate Gaussian distribution, $\Phi(\cdot)$.

Definition 9.40 (Gaussian Copula). The Gaussian copula is

$$C(u_1, u_2, \dots, u_k) = \Phi_k(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \dots, \Phi^{-1}(u_k)) \quad (9.78)$$

where $\Phi^{-1}(\cdot)$ is the inverse of the univariate Gaussian distribution function.

Recall that if U is a uniform random variable then $X = \Phi^{-1}(U)$ is distributed standard normal. This transformation allows the Gaussian copula density to be implicitly defined using the inverse distribution function. The Gaussian copula density is

$$c(u_1, u_2, \dots, u_k) = \frac{(2\pi)^{-\frac{k}{2}} |\mathbf{R}|^{-\frac{1}{2}} \exp(-\frac{1}{2}\boldsymbol{\eta}' \mathbf{R}^{-1} \boldsymbol{\eta})}{\phi(\Phi^{-1}(u_1)) \dots \phi(\Phi^{-1}(u_k))} \quad (9.79)$$

where $\boldsymbol{\eta} = \Phi^{-1}(\mathbf{u})$ is a k by 1 vector where $\eta_i = \Phi^{-1}(u_i)$, \mathbf{R} is a correlation matrix and $\phi(\cdot)$ is the normal PDF. The extra terms in the denominator are present in all implicitly defined copulas since the joint density is the product of the marginal densities and the copula density.

$$\begin{aligned} f_1(x_1) \dots f_k(x_k) c(u_1, \dots, u_k) &= f(x_1, x_2, \dots, x_k) \\ c(u_1, \dots, u_k) &= \frac{f(x_1, x_2, \dots, x_k)}{f_1(x_1) \dots f_k(x_k)} \end{aligned}$$

9.7.3.4 Student's t Copula

The Student's t copula is also implicitly defined using the multivariate Student's t distribution.

Definition 9.41 (Student's Copula). The Student's t copula is

$$C(u_1, u_2, \dots, u_k) = t_{k,v}(t_v^{-1}(u_1), t_v^{-1}(u_2), \dots, t_v^{-1}(u_k)) \quad (9.80)$$

where $t_{k,v}(\cdot)$ is the k -variate Student's t distribution function with v degrees of freedom and t_v^{-1} is the inverse of the univariate Student's t distribution function with v degrees of freedom.

Note that while the Student's t distribution is superficially similar to a normal distribution, variables that are distributed multivariate t_v are substantially more dependent if v is small (3 – 8). A multivariate Student's t is defined as a multivariate normal divided by a single, common, independent χ_v^2 standardized to have mean 1. When v is small, the chance of seeing a small value in the denominator is large, and since this divisor is common, all series tend to take relatively large values simultaneously.

9.7.3.5 Clayton Copula

The Clayton copula exhibits asymmetric dependence for most parameter values. The lower tail is more dependent than the upper tail, and so it may be appropriate for modeling the returns of some financial assets, e.g., equities.

Definition 9.42 (Clayton Copula). The Clayton copula is

$$C(u_1, u_2) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}, \theta > 0 \quad (9.81)$$

The Clayton copula limits to the independence copula when $\theta \rightarrow 0$. The copula density can be found by differentiating the Copula with respect to u_1 and u_2 , and so is

$$c(u_1, u_2) = (\theta + 1) u_1^{-\theta-1} u_2^{-\theta-1} (u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta-2}.$$

9.7.3.6 Gumbel and Rotated Gumbel Copula

The Gumbel copula exhibits asymmetric dependence in the upper tail rather than the lower tail.

Definition 9.43 (Gumbel Copula). The Gumbel copula is

$$C(u_1, u_2) = \exp \left[- \left((-\ln u_1)^\theta + (-\ln u_2)^\theta \right)^{1/\theta} \right], \theta \geq 1 \quad (9.82)$$

The Gumbel copula exhibits upper tail dependence that is increasing in θ . It approaches to the independence copula as $\theta \rightarrow 1$. Because upper tail dependence is relatively rare among financial assets, a “rotated” version of the Gumbel is more useful when modeling financial asset returns.

Let $C(u_1, u_2)$ be a bivariate copula. The rotated version¹² of the copula is given by

$$C^R(u_1, u_2) = u_1 + u_2 - 1 + C(1 - u_1, 1 - u_2).$$

¹²The rotated copula is commonly known as the survival copula, since rather than computing the probability of observing values smaller than (u_1, u_2) , it computes the probability of seeing values larger than (u_1, u_2) .

Using this definition allows the rotated Gumbel copula to capture lower tail dependence rather than upper tail dependence.

Definition 9.44 (Rotated Gumbel Copula). The rotated Gumbel copula is

$$C^R(u_1, u_2) = u_1 + u_2 - 1 + \exp \left[- \left((-\ln(1-u_1))^\theta + (-\ln(1-u_2))^\theta \right)^{1/\theta} \right], \theta \geq 1 \quad (9.83)$$

The rotated Gumbel is the Gumbel copula using $1-u_1$ and $1-u_2$ as its arguments. The extra terms are used to satisfy the four properties of a copula. The rotated Gumbel copula density is tedious to compute but is presented here.

The rotated Gumbel copula density is

$$\begin{aligned} c(u_1, u_2) &= \frac{\exp \left[- \left((-\ln(1-u_1))^\theta + (-\ln(1-u_2))^\theta \right)^{1/\theta} \right] ((-\ln(1-u_1))(-\ln(1-u_2)))^{\theta-1}}{(1-u_1)(1-u_2)((-\ln(1-u_1)) + (-\ln(1-u_2)))^{2-1/\theta}} \\ &\quad \times \left(\left((-\ln(1-u_1))^\theta + (-\ln(1-u_2))^\theta \right)^{1/\theta} + \theta - 1 \right). \end{aligned}$$

This copula density is identical to the Gumbel copula density only using $1-u_1$ and $1-u_2$ as its arguments. The rotation moves values near zero, where the dependence is low, to be near one, where the dependence is higher.

9.7.3.7 Joe-Clayton Copula

The Joe-Clayton copula allows for asymmetric dependence in both tails.

Definition 9.45 (Joe-Clayton Copula). The Joe-Clayton copula is

$$C(u_1, u_2) = 1 - \left(1 - \left[\left(1 - (1-u_1)^{\theta_U} \right)^{-\theta_L} + \left(1 - (1-u_2)^{\theta_U} \right)^{-\theta_L} - 1 \right]^{-1/\theta_L} \right)^{1/\theta_U} \quad (9.84)$$

where the two parameters, θ_L and θ_U are directly related to lower and upper tail dependence through

$$\theta_L = -\frac{1}{\log_2(\tau^L)}, \quad \theta_U = \frac{1}{\log_2(2-\tau^U)}$$

where both coefficients of tail dependence satisfy $0 < \tau^i < 1$, $i = L, U$.

Deriving the density of a Joe-Clayton copula is a straightforward, but tedious, calculation. The Joe-Clayton copula is not symmetric, even when the same values for τ^L and τ^U are used. This asymmetry may be acceptable, but if symmetry is preferred a symmetrized copula can be constructed by averaging a copula with its rotated counterpart.

Definition 9.46 (Symmetrized Copula). Let $C(u_1, u_2)$ be an asymmetric bivariate copula. The symmetrized version of the copula is given by

$$C^S(u_1, u_2) = \frac{1}{2} (C(u_1, u_2) + C^R(1-u_1, 1-u_2)) \quad (9.85)$$

If $C(u_1, u_2)$ is already symmetric, then $C(u_1, u_2) = C^R(1-u_1, 1-u_2)$ and so the $C^S(u_1, u_2)$ must also be symmetric. The copula density, assuming it exists, is

$$c^S(u_1, u_2) = \frac{1}{2} (c(u_1, u_2) + c^R(1-u_1, 1-u_2)).$$

Copula	τ^L	τ^U	Notes
Gaussian	0	0	$ \rho < 1$
Students t	$2t_{v+1}(w)$	$2t_{v+1}(w)$	$w = -\sqrt{v+1}\sqrt{1-\rho}/\sqrt{1+\rho}$
Clayton	$2^{-\frac{1}{\theta}}$	0	
Gumbel	0	$2 - 2^{\frac{1}{\theta}}$	Rotated Swaps τ^L and τ^U
Symmetrized Gumbel	$1 - 2^{\frac{1-\theta}{\theta}}$	$1 - 2^{\frac{1-\theta}{\theta}}$	
Joe-Clayton	$2^{-\frac{1}{\theta_L}}$	$2 - 2^{\frac{1}{\theta_U}}$	Also Symmetrized JC

Table 9.7: The relationship between parameter values and tail dependence for the copulas in section 9.7.3. $t_{v+1}(\cdot)$ is the CDF of a univariate Student's t distribution with $v + 1$ degree of freedom.

9.7.4 Tail Dependence in Copulas

The copulas presented in the previous section all have different functional forms, and so produce different distributions. One simple method to compare the different forms is through the tail dependence. Table 9.7 show the relationship between the tail dependence in the different copulas and their parameters. The Gaussian has no tail dependence except in the extreme case when $|\rho| = 1$, in which case tail dependence is 1 in both tails. Other copulas, such as the Clayton and Gumbel, have asymmetric tail dependence.

9.7.5 Visualizing Copulas

Copulas are defined on the unit hypercube (or unit square in a bivariate copula), and so one obvious method to inspect the difference between two is to plot the distribution function or the density on its default domain. This visualization method does not facilitate inspecting the tail dependence which occurs in the small squares of in $[0, 0.05] \times [0, 0.05]$ and $[.95, 1] \times [.95, 1]$, lower and upper 5% of each margin. Transforming the marginal distribution of each series to be standard normal is a superior method to visualize the dependence in the copula. This visualization ensures that any differences are attributable to the copula while distributing the interesting aspects over a wider range of values. It also projects the dependence structure into a familiar space that more closely resembles two financial asset returns.

Figure 9.14 contains plots of 4 copulas. The top two panels show the independence copula and the comonotonicity copula as distributions on $[0, 1] \times [0, 1]$ where curves are isoprobability lines. In distribution space, high dependence appears as an "L" shape and independence appears as a parabola. The bottom two figures contain the normal copula distribution and the Gaussian copula density using normal margins, where in both cases the correlation is $\rho = 0.5$. The Gaussian copula is more dependent than the independence copula – a special case of the Gaussian copula when $\rho = 0$ – but less dependent than the comonotonicity copula (except when $\rho = 1$). The density has both a Gaussian copula and Gaussian margins, and so depicts a bivariate normal. The density function shows the dependence between the two series in a more transparent manner.¹³

Figure 9.15 contains plots of 4 copulas depicted as densities with standard normal marginal distri-

¹³Some copulas do not have a copula density, and in these cases, the copula distribution is the only visualization option.

butions. The upper left panel contains the Clayton density which has lower-tail dependence ($\theta = 1.5$). The upper right shows the symmetrized Joe-Clayton where $\tau^L = \tau^U = 0.5$, which has both upper and lower tail dependence. The bottom two panels show the rotated Gumbel and symmetrized Gumbel where $\theta = 1.5$. The rotated Gumbel is similar to the Clayton copula although it is not identical.

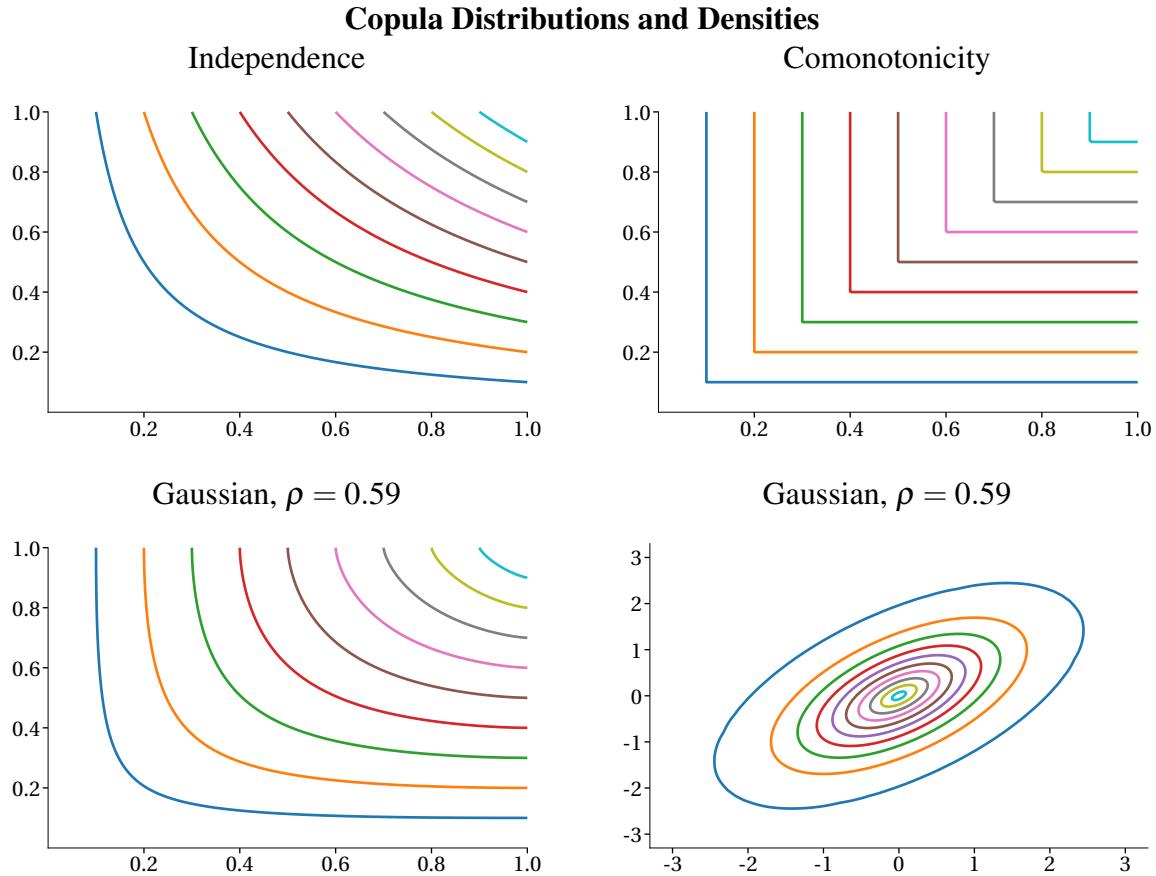


Figure 9.14: The top left panel shows the isoprobability curves of an independence copula. The top right panel shows the isoprobability curves of the comonotonicity copula, which has perfect dependence. The bottom panels contain the Gaussian copula, where the left depicts the copula in distribution space ($[0, 1] \times [0, 1]$) and the right shows the copula density using standard normal marginal distributions. The correlation of the Gaussian copula is estimated using weekly returns on the S&P 500 and FTSE 100.

9.7.6 Estimation of Copula models

A copula-based model is a joint distribution, and so parameters can be estimated using maximum likelihood. As long as the copula density exists, and the parameters of the margins are distinct from the parameters of the copula (which is almost always the case), the likelihood of a k -variate random variable Y can be written as

$$f(\mathbf{y}_t; \boldsymbol{\theta}, \psi) = f_1(y_{1,t}; \boldsymbol{\theta}_1) f_2(y_{2,t}; \boldsymbol{\theta}_2) \dots f_k(y_{k,t}; \boldsymbol{\theta}_k) c(u_{1,t}, u_{2,t}, \dots, u_{k,t}; \psi)$$

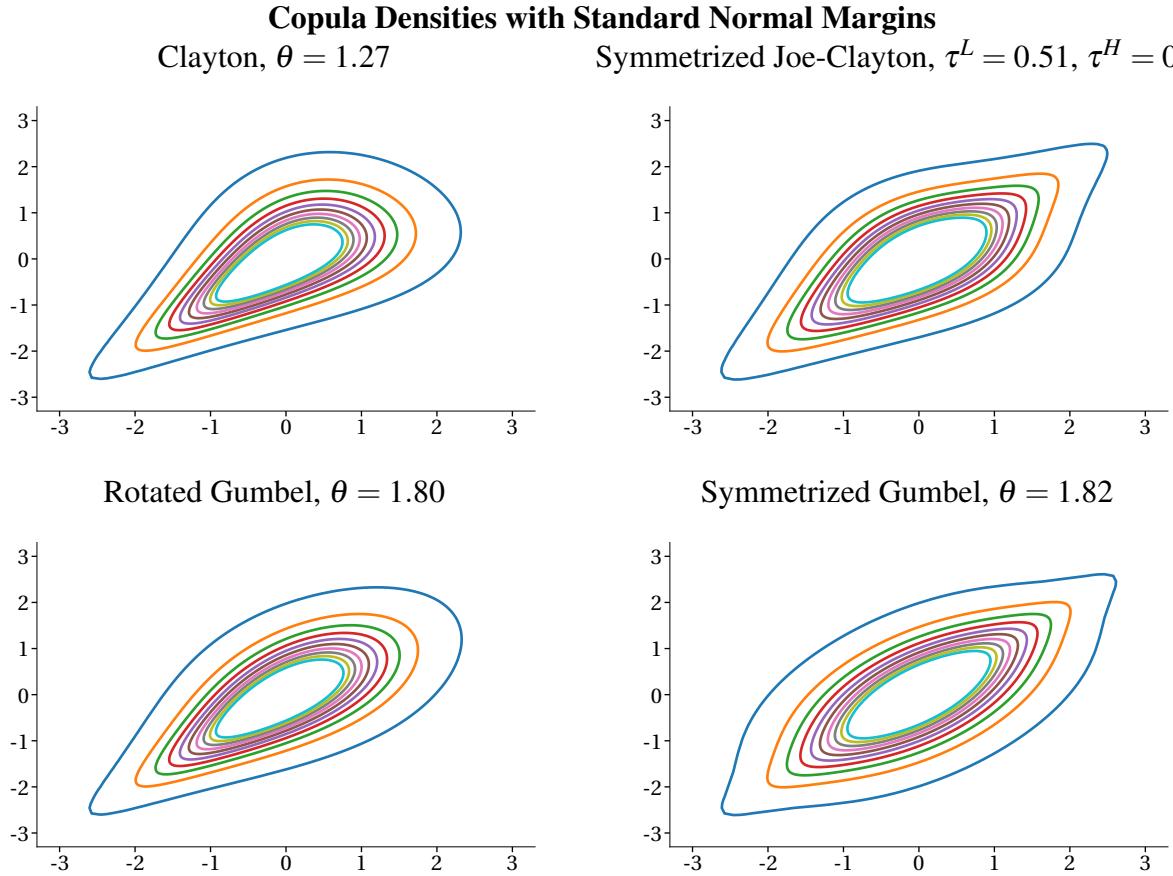


Figure 9.15: These four panels all depict copulas as densities using standard normal margins. All differences in appearance can be attributed to the differences in the copulas. The top left panel contains the Clayton copula density. The top right contains the symmetrized Joe-Clayton. The bottom panels contain the rotated Gumbel which has lower tail dependence and the symmetrized Gumbel. The parameter values are estimated from weekly returns on the S&P 500 and FTSE 100.

where $u_{j,t} = F_j^{-1}(y_{j,t}; \theta_j)$ are the probability integral transformed observations, θ_j are the parameters specific to marginal model j and ψ are the parameters of the copula. The log likelihood is then the sum of the marginal log likelihoods and the copula log likelihood,

$$l(\theta, \psi; \mathbf{y}) = \ln f_1(y_1; \theta_1) + \ln f_2(y_2; \theta_2) + \dots + \ln f_k(y_k; \theta_k) + \ln c(u_1, u_2, \dots, u_k; \psi).$$

This decomposition allows for consistent estimation of the parameters in two steps:

1. For each margin j , estimate θ_j using quasi maximum likelihood as the solution to

$$\arg \max_{\theta_j} \sum_{t=1}^T \ln f_j(y_{j,t}; \theta_j).$$

When fitting models using copulas it is also important to verify that the marginal models are adequate using the diagnostics for univariate densities described in chapter 8.

2. Using the probability integral transformed residuals evaluated at the estimated parameters, $\hat{u}_{j,t} = F^{-1}(y_{j,t}; \hat{\theta}_j)$, estimate the parameters of the copula as

$$\arg \max_{\psi} \sum_{t=1}^T \ln c(\hat{u}_{1,t}, \hat{u}_{2,t}, \dots, \hat{u}_{k,t}; \psi).$$

This two-step procedure is not efficient in the sense that the parameter estimates are consistent but have higher variance than if all parameters are simultaneously estimated. In practice, the reduction in precision is typically small. If parameter estimation accuracy is an important consideration, then the two-step estimator can be used as a starting value for an estimator which simultaneously estimates all parameters. Standard errors can be computed from the two-step estimation procedure by treating it as a two-step GMM problem where the scores of the marginal log likelihoods and the copula are the moment conditions (See section 6.10 for a discussion).

An alternative estimation procedure uses *nonparametric* models for the margins. Nonparametric margins are typically employed when characterizing the distribution of the margins are not particularly important, and so the first step can be replaced through the use of the empirical CDF. The empirical CDF estimates the $\hat{u}_{j,t} = \text{rank}(y_{j,t})/(T+1)$ where rank is the ordinal rank of observation t among the T observations. The empirical CDF is uniform by construction. Using the empirical CDF is not generally appropriate when the data have time-series dependence (e.g., volatility clustering) or when forecasting is an important consideration.

9.7.7 Application: Dependence between the S&P 500 and the FTSE 100

The use of copulas is illustrated using returns of the S&P 500 and the FTSE 100. Weekly returns are used to mitigate issues with non-synchronous closing times. The first example uses the empirical CDF to transform the returns so that the model focuses on the unconditional dependence between the two series. The upper left panel of figure 9.16 contains a scatter plot of the ECDF transformed residuals. The residuals tend to cluster around the 45° line indicating positive dependence (correlation). There are clusters of observations near $(0,0)$ and $(1,1)$ that indicating the returns have tail dependence. The normal, students t , Clayton, rotated Gumbel, symmetrized Gumbel and symmetric Joe-Clayton copulas are all estimated. Parameter estimates and copula log-likelihoods are reported in Table 9.8. The Joe-Clayton fits the data the best, followed by the symmetrized Gumbel and then the rotated Gumbel. The Clayton and the Gaussian both appear to fit the data substantially worse than the others. In the Joe-Clayton, both tails appear to have some dependence, although returns in the lower tails are substantially more dependent.

Copulas can also be used in conditional density models. Combining a constant copula with dynamic models of each margin is similar to using a CCC-GARCH model to estimate the conditional covariance. A conditional joint density model is built using TARCH(1,1,1) volatilities with skew t errors for each index return series. The copulas are estimated using the conditionally transformed residuals $\hat{u}_{i,t} = F(r_{i,t}; \hat{\sigma}_t^2, \hat{v}, \hat{\lambda})$ where σ_t^2 is the conditional variance, v is the degree of freedom, and λ captures the skewness in the standardized residuals. Parameter estimates are reported in Table 9.9. The top panel reports the parameter estimates from the TARCH model. Both series have persistent volatility although the leverage effect is stronger in the S&P 500 than it is in the FTSE 100. Standardized residuals in the S&P 500 are heavier tailed, and both are negatively skewed.

Dependence Measures for Weekly FTSE and S&P 500 Returns

Copula	θ_1	θ_2	Log. Lik.	τ^L	τ^U
Gaussian	0.645		-486.9	0	0
Clayton	1.275		-460.2	0.581	0
Rot. Gumbel	1.805		-526.2	0.532	0
Sym. Gumbel	1.828		-529.5	0.270	0.270
Sym. Joe-Clayton	0.518	0.417	-539.9	0.518	0.417

Table 9.8: Parameter estimates for the unconditional copula between weekly returns on the S&P 500 and the FTSE 100. Marginal distributions are estimated using empirical CDFs. For the Gaussian copula, θ_1 is the correlation, and in the Joe-Clayton θ_1 is τ^L and θ_2 is τ^U . The third column reports the log likelihood from the copula density. The final two columns report the estimated lower and upper tail dependence.

The parameter estimates using the conditional marginals all indicate less dependence than those estimated using the empirical CDF. This reduction in dependence is due to synchronization between the volatility of the two markets. Coordinated periods of high volatility leads to large returns in both series at the same time, even when the standardized shock is only moderately large. Unconditional models use data from both high and low volatility periods. Volatility and dependence are linked in financial markets, and so ignoring conditional information tends to higher unconditional dependence than conditional dependence. This phenomenon is similar to the generation of heavy tails in the unconditional distribution of a single asset return – mixing periods of high and low volatility produced heavy tails. The difference in the dependence shows up in the parameter values in the copulas, and in the estimated tail indices, which are uniformly smaller than in their unconditional counterparts. The changes in dependence also appear through the reduction in the improvement in the log-likelihoods of the dependent copulas relative to the Gaussian.

Figure 9.16 contains some diagnostic plots related to fitting the conditional copula. The top right panel contains the scatter plot of the probability integral transformed residuals using from the TARCH. While these appear similar to the plot from the empirical CDF, the amount of clustering near $(0, 0)$ and $(1, 1)$ is slightly lower. The bottom left panel contains a QQ plot of the actual returns against the expected returns using the estimated degree of freedom and skewness parameters. These curves are straight except for the most extreme observations, and so indicate an acceptable fit. The bottom right plot contains the annualized volatility series for the two assets where the coordination in the conditional volatilities is apparent. It also appears the coordination in volatility cycles has strengthened post-2000.

9.7.8 Dynamic Copulas

This chapter has focused on static copulas of the form $C(u_1, u_2; \theta)$. It is possible to model dependence using conditional copulas where the copula parameters evolve through time, $C(u_1, u_2; \theta_t)$, using GARCH-like dynamics. Patton (2006) first used this structure in an application to exchange rates. The primary difficulty in specifying dynamic copula models is in determining the form of the “shock”. In ARCH-type volatility models $\varepsilon_t^2 = (r_t - \mu)^2$ is the natural shock since its conditional expectation

Conditional Copula Estimates for Weekly FTSE and S&P 500 Returns

Index	α_1	γ_1	β_1	v	λ
S&P 500	0.026	0.178	0.855	8.924	-0.231
FTSE 100	0.038	0.145	0.861	8.293	-0.138
Copula	θ_1	θ_2	Log. Lik.	τ^L	τ^U
Gaussian	0.621		-439.1	0	0
Clayton	1.126		-399.6	0.540	0
Rot. Gumbel	1.713		-452.4	0.501	0
Sym. Gumbel	1.754		-452.7	0.258	0.258
Sym. Joe-Clayton	0.475	0.357	-448.5	0.475	0.357

Table 9.9: Parameter estimates for the conditional copula between weekly returns on the S&P 500 and the FTSE 100. Marginal distributions are estimated using a TARCH(1,1,1) with Hansen's Skew t error. Parameter estimates from the marginal models are reported in the top panel. The bottom panel contains parameter estimates from copulas fit using the conditionally probability integral transformed residuals. For the Gaussian copula, θ_1 is the correlation, and in the Joe-Clayton θ_1 is τ^L and θ_2 is τ^U . The third column reports the log likelihood from the copula density. The final two columns report the estimated lower and upper tail dependence.

is the variance, $E_{t-1} [\varepsilon_t^2] = \sigma_t^2$. In most copula models there is no obvious equivalent. Creal, Koopman, and Lucas (2013) have recently developed a general framework which can be used to construct a natural shock even in complex models, and have applied their methodology to estimate conditional copulas.

DCC can also be used as a dynamic Gaussian copula where the first step is modified from fitting the conditional variance to fitting the conditional distribution. Probability integral transformed residuals from the first step are then transformed to be Gaussian, and these are used to estimate the correlation parameters in the second step of the DCC estimator. The combined model has flexible marginal distributions and a Gaussian copula.

9.A Bootstrap Standard Errors

The Bootstrap is a computational tool that has a variety of uses, including estimating standard errors and simulating returns. It is particularly useful when evaluating expressions for asymptotic standard errors that are complex. This appendix provides a *very* brief introduction to bootstrap standard errors. The key intuition that underlies the bootstrap is simple. If $\{\mathbf{r}_t\}$ is a sample of T data points from some unknown joint distribution F , then $\{\mathbf{r}_t\}$ can be used to simulate (via re-sampling) from the unknown distribution F . The name bootstrap comes from the expression "To pull yourself up by your bootstraps", a seemingly impossible task, much like simulating values from an unknown distribution.

There are many implementations of the bootstrap, and each uses a different sampling scheme when generating bootstrap samples. The assumed data generating process determines which bootstraps are applicable. Bootstrap methods can be classified as parametric or non-parametric. Parametric

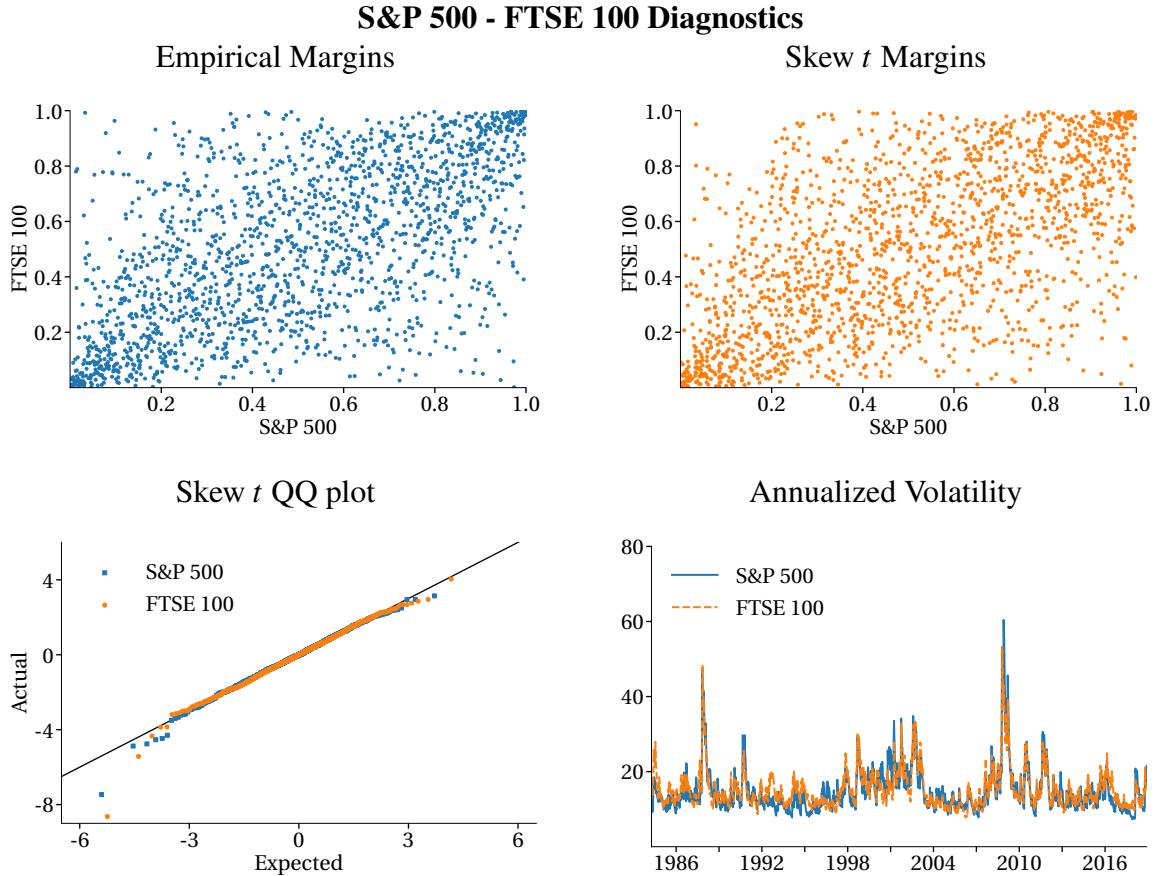


Figure 9.16: These four panels show diagnostics from fitting copulas to weekly returns on the S&P 500 and FTSE 100. The top two panels contain plots of the probability integral transformed residuals. The left panel shows the PITs constructed using the empirical CDF, and so depicts the unconditional dependence. The right contains the PITs from a TARCH(1,1,1) with Skew t errors. The bottom left contains a QQ plot of the data against the typical value from a Skew t . The bottom right plot contains the fit annualized volatility for the two indices.

bootstraps resample model residuals. Nonparametric bootstraps directly resample from the observed data and so do not rely on a model. In many applications both types of bootstraps are valid, and the choice between the two is similar to the choice between parametric and non-parametric estimators: parametric estimators are precise but may be misleading if the model is misspecified while non-parametric estimators are consistent but may require larger samples to be reliable. This appendix describes three bootstraps and one method to compute standard errors using a nonparametric bootstrap method. Comprehensive treatments of the bootstrap can be found in Efron and Tibshirani (1998) and Chernick (2008).

The i.i.d. bootstrap uses the simplest sampling scheme and is applicable when the data are i.i.d., or more generally when the model errors are not serially correlated.¹⁴

¹⁴The definition of the model error depends on the statistic of interest. For example, when bootstrapping the sample mean, the i.i.d. bootstrap can be used if the data are serially uncorrelated. When bootstrapping a variance estimator, the squared deviations must be uncorrelated. In applications of ML, the scores from the model must be serially uncorrelated.

Algorithm 9.1 (IID Bootstrap). 1. Draw T indices $\tau_i = \lceil Tu_i \rceil$ where $u_i \stackrel{i.i.d.}{\sim} U(0, 1)$ and $\lceil \cdot \rceil$ is the ceiling operator.

2. Construct an artificial time series using the indices $\{\tau_i\}_{i=1}^T$,

$$y_{\tau_1} y_{\tau_2} \dots y_{\tau_T}.$$

3. Repeat steps 1–2 a total of B times.

It is implausible to assume that the data are i.i.d. in most applications in finance, and so a bootstrap designed for dependent data is required. The two most common bootstraps for dependent data are the Circular Block Bootstrap and the Stationary Bootstrap (Politis and Romano, 1994). The Circular Block Bootstrap is based on the idea of drawing blocks of data which are sufficiently long so that the blocks are approximately i.i.d.

Algorithm 9.2 (Circular Block Bootstrap).

1. Draw $\tau_1 = \lceil Tu \rceil$ where $u \stackrel{i.i.d.}{\sim} U(0, 1)$.

2. For $i = 2, \dots, T$, if $i \bmod m \neq 0$, $\tau_i = \tau_{i-1} + 1$ where wrapping is used so that if $\tau_{i-1} = T$ then $\tau_i = 1$. If $i \bmod m = 0$ when $\tau_i = \lceil Tu \rceil$ where $u \stackrel{i.i.d.}{\sim} U(0, 1)$.

3. Construct an artificial time series using the indices $\{\tau_i\}_{i=1}^T$.

4. Repeat steps 1 – 3 a total of B times.

The Stationary Bootstrap is closely related to the block bootstrap. The only difference is that it uses blocks with lengths that are exponentially distributed with an average length of m .

Algorithm 9.3 (Stationary Bootstrap).

1. Draw $\tau_1 = \lceil Tu \rceil$ where $u \stackrel{i.i.d.}{\sim} U(0, 1)$.

2. For $i = 2, \dots, T$, draw a standard uniform $v \stackrel{i.i.d.}{\sim} U(0, 1)$. If $v > 1/m$, $\tau_i = \tau_{i-1} + 1$, where wrapping is used so that if $\tau_{i-1} = T$ then $\tau_i = 1$. If $v \leq 1/m$, $\tau_i = \lceil Tu \rceil$ where $u \stackrel{i.i.d.}{\sim} U(0, 1)$

3. Construct an artificial time series using the indices $\{\tau_i\}_{i=1}^T$.

4. Repeat steps 1 – 3 a total of B times.

In both the Circular Block Bootstrap and the Stationary Bootstrap, the block length should be chosen to capture most of the dependence in the data. The block size should not be larger than \sqrt{T} . Patton, Politis, and White (2009) provide a data-based method to select the block size in these bootstraps.

The re-sampled data are then used to make inference on statistics of interest.

Additionally, when using a nonparametric bootstrap, the i.i.d. bootstrap is only applicable when the model does not impose a time-series structure (i.e., the model is not an ARMA or GARCH).

Algorithm 9.4 (Bootstrap Parameter Covariance Estimation). *1. Begin by computing the statistic of interest $\hat{\theta}$ using the original sample.*

2. *Using a bootstrap appropriate for the dependence in the data, estimate the statistic of interest on the B artificial samples, and denote these estimates as $\tilde{\theta}_j$, $j = 1, 2, \dots, B$.*
3. *Construct confidence intervals using:*

- (a) *(Inference using standard deviation) Estimate the variance of $\hat{\theta} - \theta_0$ as*

$$B^{-1} \sum_{b=1}^B (\tilde{\theta}_b - \hat{\theta})^2$$

- (b) *(Inference using symmetric quantiles) Construct bootstrap errors as $\eta_b = \tilde{\theta}_b - \hat{\theta}$, and construct the $1 - \alpha$ confidence interval $(\hat{\theta} \pm \bar{q}_{\alpha/2})$ using the $1 - \alpha/2$ quantile of $|\eta_b|$, denoted $\bar{q}_{1-\alpha/2}$.*
- (c) *(Inference using asymmetric quantiles) Construct bootstrap errors as $\eta_b = \tilde{\theta}_b - \hat{\theta}$, and construct the $1 - \alpha$ confidence interval $(\hat{\theta} - q_{\alpha/2}, \hat{\theta} + q_{1-\alpha/2})$ using the $\alpha/2$ and $1 - \alpha/2$ quantile of η_b , denoted $q_{\alpha/2}$ and $q_{1-\alpha/2}$, respectively. The confidence interval can be equivalently defined as $(\tilde{q}_{\alpha/2}, \tilde{q}_{1-\alpha/2})$ where $\tilde{q}_{\alpha/2}$ is the $\alpha/2$ quantile of the estimators computed from the bootstrap samples, $\{\tilde{\theta}_j\}_{j=1}^B$, and $\tilde{q}_{1-\alpha/2}$ is similarly defined using the $1 - \alpha/2$ quantile.*

The bootstrap confidence intervals in this chapter are all computed using this algorithm and a stationary bootstrap with $m \propto \sqrt{T}$.

Warning: The bootstrap is broadly applicable in cases where parameters are asymptotically normal such as in regression with stationary data. They are either not appropriate or require special construction in many situations where estimators have non-standard distributions, e.g., unit roots, and so before computing bootstrap standard errors, it is useful to verify that the bootstrap produces valid inference. In cases where the bootstrap fails, subsampling, a more general statistical technique can be used to make correct inference.

Shorter Problems

Problem 9.1. Describe the observable factor covariance model and the exponentially weighted moving average covariance model. Discuss the relative strengths and weaknesses of these two models.

Problem 9.2. Describe one multivariate GARCH model and one multivariate volatility model which is not a GARCH specification. Describe the relative strengths and weaknesses of these two models.

Problem 9.3. Discuss three alternative models for conditional covariance.

Problem 9.4. What is Exceedance Correlation?

Problem 9.5. Compare and contrast linear and rank correlation.

Longer Questions

Exercise 9.1. Answer the following questions about covariance modeling

1. Describe the similarities between the RiskMetrics 1994 and RiskMetrics 2006 methodologies.
2. Describe two multivariate GARCH models. What are the strengths and weaknesses of these models?
3. Other than linear correlation, describe two other measures of dependence.
4. What is Realized Covariance?
5. What are the important considerations when estimating covariance using Realized Covariance?

Exercise 9.2. Answer the following questions.

1. Briefly outline two applications in finance where a multivariate volatility models are useful.
2. Describe two of the main problems faced in multivariate volatility modeling, using two different models to illustrate these problems.
3. Recall that, in a bivariate application, the BEKK model of a time-varying conditional covariance matrix is:

$$\begin{bmatrix} \sigma_{11,t} & \sigma_{12,t} \\ \sigma_{12,t} & \sigma_{22,t} \end{bmatrix} \equiv \Sigma_t = \mathbf{C}\mathbf{C}' + \mathbf{A}\boldsymbol{\varepsilon}_{t-1}\boldsymbol{\varepsilon}'_{t-1}\mathbf{A}' + \mathbf{B}\Sigma_{t-1}\mathbf{B}'$$

where \mathbf{C} is a lower triangular matrix, and $\boldsymbol{\varepsilon}'_t \equiv [\varepsilon_{1,t}, \varepsilon_{2,t}]$ is the vector of residuals. Using the result that $\text{vec}(\mathbf{QRS}) = (\mathbf{S}' \otimes \mathbf{Q}) \text{vec}(\mathbf{R})$, where \otimes is the Kronecker product, re-write the BEKK model for $\text{vec}(\Sigma_t)$ rather than Σ_t .

4. Estimating this model on two-day returns on the S&P 500 index and the FTSE 100 index over the period 4 April 1984 to 30 December 2008, we find

$$\hat{\mathbf{C}} = \begin{bmatrix} 0.15 & 0 \\ 0.19 & 0.20 \end{bmatrix}, \quad \hat{\mathbf{B}} = \begin{bmatrix} 0.97 & -0.01 \\ -0.01 & 0.92 \end{bmatrix}, \quad \hat{\mathbf{A}} = \begin{bmatrix} 0.25 & 0.03 \\ 0.05 & 0.32 \end{bmatrix}.$$

Using your answer from (c), compute the (1, 1) element of the coefficient matrix on $\text{vec}(\Sigma_{t-1})$.

Exercise 9.3. Answer the following questions.

1. For a set of two asset returns, recall that the BEKK model for a time-varying conditional covariance matrix is:

$$\begin{bmatrix} h_{11t} & h_{12t} \\ h_{12t} & h_{22t} \end{bmatrix} \equiv \Sigma_t = \mathbf{C}\mathbf{C}' + \mathbf{B}\Sigma_{t-1}\mathbf{B}' + \mathbf{A}\boldsymbol{\varepsilon}_{t-1}\boldsymbol{\varepsilon}'_{t-1}\mathbf{A}'$$

where \mathbf{C} is a lower triangular matrix, and $\boldsymbol{\varepsilon}'_t \equiv [\varepsilon_{1,t}, \varepsilon_{2,t}]$ is the vector of residuals.

2. Describe two of the main problems faced in multivariate volatility modeling, and how the BEKK model overcomes or does not overcome these problems.

3. Using the result that $\text{vec}(\mathbf{QRS}) = (\mathbf{S}' \otimes \mathbf{Q}) \text{vec}(\mathbf{R})$, where \otimes is the Kronecker product, re-write the BEKK model for $\text{vec}(\Sigma_t)$ rather than Σ_t .
4. Estimating this model on two-day returns on the S&P 500 index and the FTSE 100 index over the period 4 April 1984 to 30 December 2008, we find

$$\hat{\mathbf{C}} = \begin{bmatrix} 0.15 & 0 \\ 0.19 & 0.20 \end{bmatrix}, \quad \hat{\mathbf{B}} = \begin{bmatrix} 0.97 & -0.01 \\ -0.01 & 0.92 \end{bmatrix}, \quad \hat{\mathbf{A}} = \begin{bmatrix} 0.25 & 0.03 \\ 0.05 & 0.32 \end{bmatrix}.$$

Using your answer from (b), compute the estimated intercept vector in the $\text{vec}(\Sigma_t)$ representation of the BEKK model. (Hint: this vector is 4×1 .)

5. Computing “exceedance correlations” on the two-day returns on the S&P 500 index and the FTSE 100 index, we obtain Figure 9.17. Describe what exceedance correlations are, and what feature(s) of the data they are designed to measure.
6. What does the figure tell us about the dependence between returns on the S&P 500 index and returns on the FTSE 100 index?

Exercise 9.4. Answer the following questions about covariance modeling:

1. Describe the RiskMetrics 1994 methodology for modeling the conditional covariance.
2. How does the RiskMetrics 2006 methodology differ from the 1994 methodology for modeling the conditional covariance?
3. Describe one multivariate GARCH model. What are the strengths and weaknesses of the model?
4. How is the 5% portfolio *VaR* computed when using the RiskMetrics 1994 methodology?
5. Other than linear correlation, describe two measures of dependence.
6. What is Realized Covariance?
7. What are the important considerations when estimating covariance using Realized Covariance?

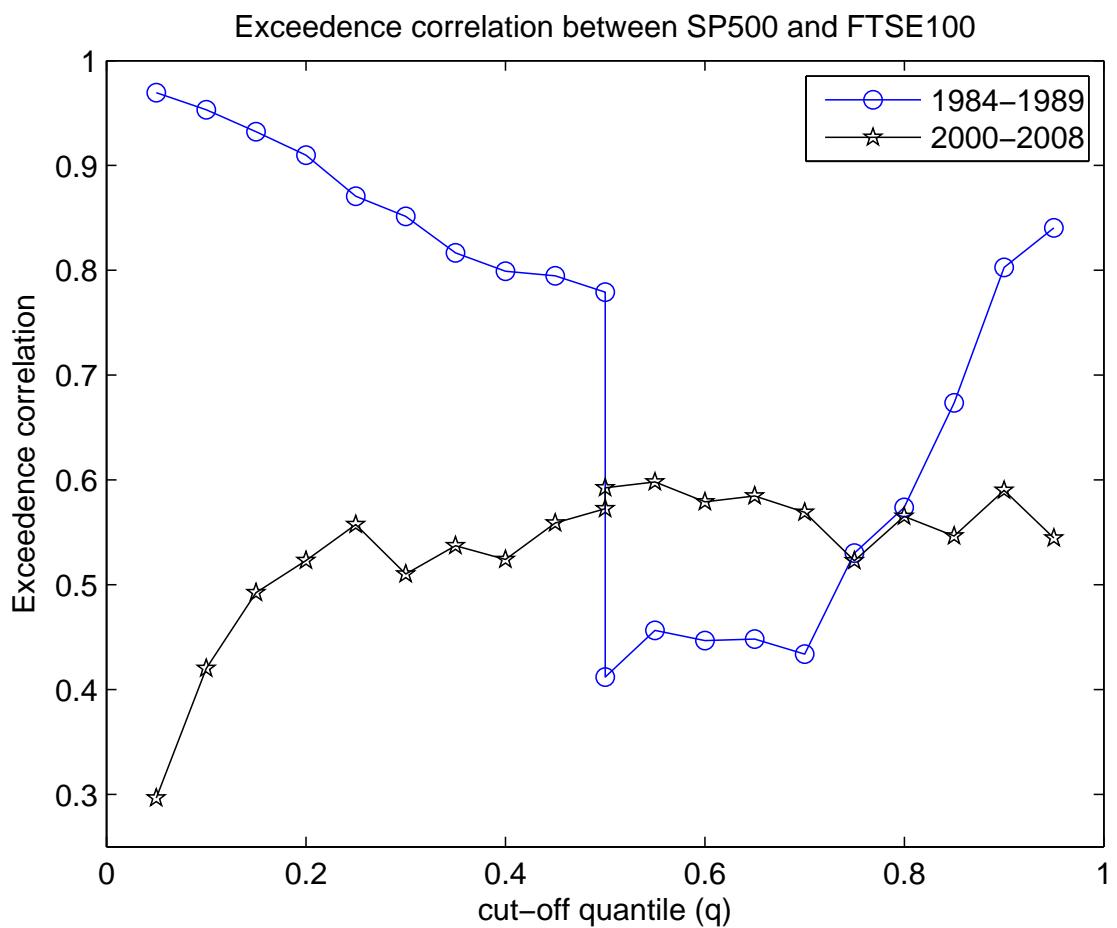


Figure 9.17: Exceedance correlations between two-day returns on the S&P 500 index and the FTSE 100 index. Line with circles uses data from April 1984 to December 1989; line with stars uses data from January 2000 to December 2008.

Bibliography

- Abramowitz, Milton and Irene Stegun (1964). *Handbook of Mathematical Functions with Forumula, Graphs and Mathematical Tables*. Dover Publications.
- Aït-Sahalia, Yacine and Andrew W. Lo (Apr. 1998). “Nonparametric Estimation of State-Price Densities Implicit in Financial Asset Prices.” In: *Journal of Finance* 53.2, pp. 499–547.
- Andersen, Torben G. and Tim Bollerslev (Nov. 1998). “Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts.” In: *International Economic Review* 39.4, pp. 885–905.
- Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Paul Labys (2003). “Modeling and Forecasting Realized Volatility.” In: *Econometrica* 71.1, pp. 3–29.
- Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Clara Vega (2007). “Real-time price discovery in global stock, bond and foreign exchange markets.” In: *Journal of International Economics* 73.2, pp. 251–277.
- Andrews, Donald W K and Werner Ploberger (1994). “Optimal Tests when a Nuisance Parameter is Present only under the Alternative.” In: *Econometrica* 62.6, pp. 1383–1414.
- Ang, Andrew, Joseph Chen, and Yuhang Xing (2006). “Downside Risk.” In: *Review of Financial Studies* 19.4, pp. 1191–1239.
- Bai, Jushan and Serena Ng (2002). “Determining the number of factors in approximate factor models.” In: *Econometrica* 70.1, pp. 191–221.
- Baldessari, Bruno (1967). “The Distribution of a Quadratic Form of Normal Random Variables.” In: *The Annals of Mathematical Statistics* 38.6, pp. 1700–1704.
- Bandi, Federico and Jeffrey Russell (2008). “Microstructure Noise, Realized Variance, and Optimal Sampling.” In: *The Review of Economic Studies* 75.2, pp. 339–369.
- Barndorff-Nielsen, Ole E., Peter Reinhard Hansen, et al. (Nov. 2008). “Designing Realized Kernels to Measure the ex-post Variation of Equity Prices in the Presence of Noise.” In: *Econometrica* 76.6, pp. 1481–1536.
- (2011). “Multivariate Realised Kernels: Consistent Positive Semi-definite Estimators of the Co-variation of Equity Prices with Noise and Non-synchronous Trading.” In: *Journal of Econometrics* 162, pp. 149–169.
- Barndorff-Nielsen, Ole E. and Neil Shephard (2004). “Econometric Analysis of Realized Covariation: High Frequency Based Covariance, Regression, and Correlation in Financial Economics.” In: *Econometrica* 72.3, pp. 885–925.
- Baxter, Marianne and Robert G King (Nov. 1999). “Measuring Business Cycles: Approximate Band-Pass Filters For Economic Time Series.” In: *The Review of Economics and Statistics* 81.4, pp. 575–593.
- Bekaert, Geert and Guojun Wu (2000). “Asymmetric Volatility and Risk in Equity Markets.” In: *The Review of Financial Studies* 13.1, pp. 1–42. DOI: [10.1093/rfs/13.1.1](https://doi.org/10.1093/rfs/13.1.1). eprint: [/oup/rfs.13.1.1](https://oup.rfs.oxfordjournals.org/content/13/1/1)

- backfile/content_public/journal/rfs/13/1/10.1093/rfs/13.1.1/2/130001.pdf. URL: <http://dx.doi.org/10.1093/rfs/13.1.1>.
- Berkowitz, Jeremy (2001). "Testing density forecasts, with applications to risk management." In: *Journal of Business & Economic Statistics* 19, pp. 465–474.
- Beveridge, S and C R Nelson (1981). "A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the 'business cycle'." In: *Journal of Monetary Economics* 7.2, pp. 151–174.
- Black, Fischer (1972). "Capital Market Equilibrium with Restricted Borrowing." In: *The Journal of Business* 45.3, pp. 444–455. ISSN: 0021-9398.
- Bollerslev, Tim (1986). "Generalized Autoregressive Conditional Heteroskedasticity." In: *Journal of Econometrics* 31.3, pp. 307–327.
- (Aug. 1987). "A Conditionally Heteroskedastic Time Series Model for Security Prices and Rates of Return Data." In: *Review of Economics and Statistics* 69.3, pp. 542–547.
- (1990). "Modeling the Coherence in Short-Run Nominal Exchange Rates: A Multivariate Generalized ARCH Model." In: *Review of Economics and Statistics* 72.3, pp. 498–505.
- Bollerslev, Tim, Robert F. Engle, and Jeffrey M. Wooldridge (1988). "A Capital Asset Pricing Model with Time-Varying Covariances." In: *Journal of Political Economy* 96.1, pp. 116–131.
- Bollerslev, Tim and Jeffrey M. Wooldridge (1992). "Quasi-Maximum Likelihood Estimation and Inference in Dynamic Models with Time-Varying Covariances." In: *Econometric Reviews* 11.2, pp. 143–172.
- Breeden, Douglas T. and Robert H. Litzenberger (1978). "Prices of State Contingent Claims Implicit in Option Prices." In: *Journal of Business* 51, pp. 621–651.
- Britten-Jones, Mark and Anthony Neuberger (2000). "Option Prices, Implied Price Processes, and Stochastic Volatility." In: *Journal of Finance* 55.2, pp. 839–866.
- Burns, Patrick, Robert F. Engle, and Joe Mezrich (1998). "Correlations and Volatilities of Asynchronous Data." In: *Journal of Derivatives* 5.4, pp. 7–18.
- Campbell, John Y. (1996). "Understanding Risk and Return." In: *Journal of Political Economy* 104, pp. 298–345.
- Cappiello, Lorenzo, Robert F. Engle, and Kevin Sheppard (2006). "Asymmetric dynamics in the correlations of global equity and bond returns." In: *Journal of Financial Econometrics* 4.4, pp. 537–572.
- Casella, George and Roger L Berger (2001). *Statistical Inference (Hardcover)*. 2nd ed. Duxbury.
- CBOE (2003). VIX: CBOE Volatility Index. Tech. rep. Chicago Board Options Exchange. URL: <http://www.cboe.com/micro/vix/vixwhite.pdf>.
- Chernick, Michael R (2008). *Bootstrap Methods: A Guide for Practitioners and Researchers*. Wiley Series in Probability and Statistics. Hoboken, New Jersey: John Wiley & Sons Inc, p. 592.
- Christie, Andrew A. (1982). "The stochastic behavior of common stock variances: Value, leverage and interest rate effects." In: *Journal of Financial Economics* 10.4, pp. 407–432. ISSN: 0304-405X. DOI: [https://doi.org/10.1016/0304-405X\(82\)90018-6](https://doi.org/10.1016/0304-405X(82)90018-6). URL: <http://www.sciencedirect.com/science/article/pii/0304405X82900186>.
- Christoffersen, Peter F (2003). *Elements of Financial Risk Management*. London: Academic Press Inc.
- Cochrane, John H (2001). *Asset Pricing*. Princeton, N. J.: Princeton University Press.
- Collard, Fabrice et al. (2018). "Ambiguity And The Historical Equity Premium." In: *Quantitative Economics* 9.2, pp. 945–993.

- Corsi, Fulvio (2009). "A Simple Approximate Long-Memory Model of Realized Volatility." In: *Journal of Financial Econometrics* 7.2, pp. 174–196. eprint: <http://jfec.oxfordjournals.org/cgi/reprint/7/2/174.pdf>.
- Creal, Drew, Siem Jan Koopman, and André Lucas (2013). "Generalized autoregressive score models with applications." In: *Journal of Applied Econometrics* 28.5, pp. 777–795.
- Davidson, Russell and James G MacKinnon (2003). *Econometric Theory and Methods*. Oxford University Press.
- Demeterfi, Kresimir et al. (1999). *More Than You Ever Wanted To Know About Volatility Swaps*. Quantitative Strategies Research Notes. Goldman Sachs.
- Diebold, Francis X. and Roberto S Mariano (July 1995). "Comparing Predictive Accuracy." In: *Journal of Business & Economic Statistics* 13.3, pp. 253–263.
- Ding, Z and R Engle (2001). "Large Scale Conditional Matrix Modeling, Estimation and Testing." In: *Academia Economic Papers* 29.2, pp. 157–184.
- Ding, Zhuanxin, Clive W. J. Granger, and Robert F. Engle (1993). "A Long Memory Property of Stock Market Returns and a New Model." In: *Journal of Empirical Finance* 1.1, pp. 83–106.
- Dittmar, Robert F (2002). "Nonlinear Pricing Kernels, Kurtosis Preference, and the Cross-Section of Equity Returns." In: *Journal of Finance* 57.1, pp. 369–403.
- Efron, Bradley, Trevor Hastie, et al. (2004). "Least angle regression." In: *Annals of Statistics* 32, pp. 407–499.
- Efron, Bradley and Robert J. Tibshirani (1998). *An introduction to the bootstrap*. Boca Raton ; London: Chapman & Hal.
- Elliott, Graham, Ted Rothenberg, and James Stock (1996). "Efficient Tests for an Autoregressive Unit Root." In: *Econometrica* 64, pp. 813–836.
- Enders, Walter (2004). *Applied econometric time series*. 2nd. Hoboken, NJ: J. Wiley.
- Engle, Robert F. (1982). "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of U.K. Inflation." In: *Econometrica* 50.4, pp. 987–1008.
- (1995). *ARCH: selected readings*. Oxford University Press, USA. ISBN: 019877432X.
- (2002a). "New Frontiers for ARCH models." In: *Journal of Applied Econometrics* 17, pp. 425–446.
- (2002b). "New frontiers for ARCH models." In: *Journal of Applied Econometrics* 17.5, pp. 425–446.
- Engle, Robert F. and Kenneth F Kroner (1995). "Multivariate Simultaneous Generalized ARCH." In: *Econometric Theory* 11.1, pp. 122–150.
- Engle, Robert F. and Li Li (1998). "Macroeconomic Announcements and Volatility of Treasury Futures."
- Engle, Robert F., David M Lilien, and Russell P Robins (Mar. 1987). "Estimating Time-varying Risk Premia in the Term Structure: The ARCH-M Model." In: *Econometrica* 55.2, pp. 391–407.
- Engle, Robert F. and Simone Manganelli (2004). "CAViaR: conditional autoregressive value at risk by regression quantiles." In: *Journal of Business & Economic Statistics* 22, pp. 367–381.
- Epps, Thomas W (1979). "Comovements in Stock Prices in the Very Short Run." In: *Journal of the American Statistical Society* 74, pp. 291–296.
- Fama, E F and J D MacBeth (1973). "Risk, Return, and Equilibrium: Empirical Tests." In: *The Journal of Political Economy* 81.3, pp. 607–636.
- Fama, Eugene F and Kenneth R French (1992). "The Cross-Section of Expected Stock Returns." In: *Journal of Finance* 47, pp. 427–465.

- Fama, Eugene F and Kenneth R French (1993). "Common Risk Factors in the returns on Stocks and Bonds." In: *Journal of Financial Economics* 33, pp. 3–56.
- Fan, Jianqing and Qiwei Yao (Aug. 2005). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer Series in Statistics. Springer, p. 552.
- Gallant, A Ronald (1997). *An Introduction to Econometric Theory: Measure-Theoretic Probability and Statistics with Applications to Economics*. Princeton University Press.
- Glosten, Lawrence R., Ravi Jagannathan, and David E. Runkle (1993). "On the Relationship between the expected value and the volatility of the nominal excess return on stocks." In: *Journal of Finance* 48.5.
- Gourieroux, Christian and Joann Jasiak (Aug. 2009). "Value at Risk." In: *Handbook of Financial Econometrics*. Ed. by Yacine Aït-Sahalia and Lars Peter Hansen. Elsevier Science.
- Greene, William H (Aug. 2007). *Econometric Analysis*. 6th ed. Prentice Hall, p. 1216.
- Grimmett, Geoffrey and David Stirzaker (2001). *Probability and Random Processes*. Oxford University Press.
- Haan, Wouter den and Andrew T Levin (2000). *Robust Covariance Estimation with Data-Dependent VAR Prewhitening Order*. Tech. rep. University of California – San Diego.
- Hall, Alastair R (2005). *Generalized Method of Moments*. Oxford: Oxford University Press.
- Hamilton, James D. (Mar. 1989). "A New Approach to Economic Analysis of Nonstationary Time Series." In: *Econometrica* 57.2, pp. 357–384.
- (1994). *Time series analysis*. Princeton, N.J.: Princeton University Press.
- Hannan, E. J. and B. G. Quinn (1979). "The Determination of the Order of an Autoregression." In: *Journal of the Royal Statistical Society. Series B (Methodological)* 41.2, pp. 190–195. ISSN: 0035-9246. URL: <http://www.jstor.org/stable/2985032>.
- Hansen, Bruce E. (Aug. 1994). "Autoregressive Conditional Density Estimation." In: *International Economic Review* 35.3, pp. 705–730.
- Hansen, Lars Peter (1982). "Large Sample Properties of Generalized Method of Moments Estimators." In: *Econometrica* 50.4, pp. 1029–1054.
- Hansen, Lars Peter and Kenneth J Singleton (1982). "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models." In: *Econometrica* 50.5, pp. 1269–1286.
- Hansen, Peter Reinhard and Asger Lunde (2005). "A Realized Variance for the Whole Day Based on Intermittent High-Frequency Data." In: *Journal of Financial Econometrics* 3.4, pp. 525–554. DOI: [10.1093/jjfinec/nbi028](https://doi.org/10.1093/jjfinec/nbi028). URL: <http://jfec.oxfordjournals.org/cgi/content/abstract/3/4/525>.
- (2006). "Consistent ranking of volatility models." In: *Journal of Econometrics* 127.1-2, pp. 97–121.
- Hastie, Trevor et al. (2007). "Forward stagewise regression and the monotone lasso." In: *Electronic Journal of Statistics* 1.1, pp. 1–29.
- Hayashi, Fumio (2000). *Econometrics*. Princeton University Press.
- Hodrick, Robert J and Edward C Prescott (Feb. 1997). "Postwar U.S. Business Cycles: An Empirical Investigation." In: *Journal of Money, Credit and Banking* 29.1, pp. 1–16.
- Hong, Yongmiao, Jun Tu, and Guofu Zhou (2007). "Asymmetries in Stock Returns: Statistical Tests and Economic Evaluation." In: *Rev. Financ. Stud.* 20.5, pp. 1547–1581. DOI: [10.1093/rfs/hhl037](https://doi.org/10.1093/rfs/hhl037).
- Huber, Peter J (2004). *Robust Statistics*. Hoboken, New Jersey: John Wiley & Sons Inc.

- Ivanov, Venzislav and Lutz Kilian (2005). "A Practitioner's Guide to Lag Order Selection For VAR Impulse Response Analysis." In: *Studies in Nonlinear Dynamics & Econometrics* 9.1, pp. 1219–1253. URL: <http://www.bepress.com/snde/vol9/iss1/art2>.
- J.P.Morgan/Reuters (Dec. 1996). *RiskMetrics - Technical Document*. Tech. rep. Morgan Guaranty Trust Company.
- Jiang, George J. and Yisong S. Tian (2005). "The Model-Free Implied Volatility and Its Information Content." In: *Review of Financial Studies* 18.4, pp. 1305–1342.
- Krolzig, Hans-Martin (Aug. 1997). *Markov-Switching Vector Autoregressions: Modelling, Statistical Inference, and Application to Business Cycle Analysis*. Lecture Notes in Economics and Mathematical Systems. Berlin: Springer-Verlag, p. 357.
- Lettau, Martin and Sydney Ludvigson (2001a). "Consumption, Aggregate Wealth, and Expected Stock Returns." In: *Journal of Finance* 56.8, pp. 815–849.
- (2001b). "Resurrecting the (C)CAPM: A Cross-sectional Test when Risk Premia are Time-varying." In: *Journal of Political Economy* 109.6, pp. 1238–1287.
- Lintner, John (1965). "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets." In: *The Review of Economics and Statistics* 47.1, pp. 13–37. ISSN: 0034-6535.
- Lucas, Robert E (Nov. 1978). "Asset Prices in an Exchange Economy." In: *Econometrica* 46.6, pp. 1429–1445.
- Markowitz, Harry (1959). *Portfolio Selection: Efficient Diversification of Investments*. John Wiley.
- McNeil, Alexander J, Rüdiger Frey, and Paul Embrechts (2005). *Quantitative Risk Management : Concepts, Techniques, and Tools*. Woodstock, Oxfordshire: Princeton University Press.
- Merton, Robert C (1973). "An Intertemporal Capital Asset Pricing Model." In: *Econometrica* 41, pp. 867–887.
- Mittelhammer, Ron C (1999). *Mathematical Statistics for Economics and Business*. Springer, p. 723.
- Nelson, Daniel B. (1991). "Conditional Heteroskedasticity in Asset Returns: A new approach." In: *Econometrica* 59.2, pp. 347–370.
- Nelson, Daniel B. and Charles Q. Cao (1992). "Inequality Constraints in the univariate GARCH model." In: *Journal of Business & Economic Statistics* 10.2, pp. 229–235.
- Newey, Whitney K and Kenneth D West (1987). "A Simple, Positive Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." In: *Econometrica* 55.3, pp. 703–708.
- Noureldin, Diaa, Neil Shephard, and Kevin Sheppard (2012). "Multivariate High-Frequency-Based Volatility (HEAVY) Models." In: *Journal of Applied Econometrics* 27.6, pp. 907–933.
- Patton, Andrew (2006). "Modelling Asymmetric Exchange Rate Dependence." In: *International Economic Review* 47.2, pp. 527–556.
- Patton, Andrew, Dimitris N Politis, and Halbert White (2009). "Correction to Automatic Block-Length Selection for the Dependent Bootstrap by D. Politis and H. White." In: *Econometric Reviews* 28.4, pp. 372–375. DOI: [10.1080/07474930802459016](https://doi.org/10.1080/07474930802459016).
- Patton, Andrew and Kevin Sheppard (Dec. 2009). "Evaluating Volatility Forecasts." In: *Handbook of Financial Time Series*. Ed. by Torben G. Gustav Andersen et al. Springer, p. 750.
- Perez-Quiros, Gabriel and Allan Timmermann (2000). "Firm Size and Cyclical Variations in Stock Returns." In: *Journal of Finance* 55.3, pp. 1229–1262.
- Pesaran, H. Hashem and Yongcheol Shin (Jan. 15, 1998). "Generalized impulse response analysis in linear multivariate models." In: *Economics Letters* 58.1, pp. 17–29. ISSN: 0165-1765. DOI:

- [10.1016/S0165-1765\(97\)00214-0](https://doi.org/10.1016/S0165-1765(97)00214-0). URL: <http://www.sciencedirect.com/science/article/pii/S0165176597002140>.
- Politis, D. N. and J. P. Romano (1994). “The stationary bootstrap.” In: *Journal of the American Statistical Association* 89.428, pp. 1303–1313.
- Roll, Richard (1977). “A critique of the asset pricing theory’s tests; Part I: On past and potential testability of the theory.” In: *Journal of Financial Economics* 4, pp. 129–176.
- Ross, Stephen A (1976). “The Arbitrage Theory of Capital Asset Pricing.” In: *Journal of Economic Theory* 13.3, pp. 341–360.
- Rousseeuw, Peter J and Annick M Leroy (2003). *Robust Regression and Outlier Detection*. Hoboken, New Jersey: John Wiley & Sons Inc.
- Shanken, Jay (1992). “On the Estimation of Beta-Pricing Models.” In: *The Review of Financial Studies* 5.1, pp. 1–33.
- Sharpe, William (1964). “Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk.” In: *Journal of Finance* 19, pp. 425–442.
- Sharpe, William F (1994). “The Sharpe Ratio.” In: *The Journal of Portfolio Management* 21.1, pp. 49–58.
- Sims, Christopher A. (1980). “Macroeconomics and Reality.” In: *Econometrica* 48, pp. 1–48.
- Sklar, A (1959). “Fonctions de répartition à n dimensions et leurs marges.” In: *Publ. Inst. Statist. Univ. Paris* 8, pp. 229–231.
- Taylor, Stephen J. (1986). *Modeling Financial Time Series*. John Wiley and Sons Ltd.
- (2005). *Asset price dynamics, volatility, and prediction*. Princeton, N.J.; Oxford: Princeton University Press.
- Tibshirani, Robert (1996). “Regression shrinkage and selection via the lasso.” In: *Journal of the Royal Statistical Society, Series B* 58, pp. 267–288.
- Tong, H (1978). “On a threshold model.” In: *Pattern Recognition and Signal Processing*. Ed. by C H Chen. Amsterdam: Sijhoff and Noordhoff.
- Veronesi, Pietro (1999). “Stock Market Overreaction to Bad News in Good Times: A Rational Expectations Equilibrium Model.” In: *The Review of Financial Studies* 12, pp. 975–1007.
- Wackerly, Dennis, William Mendenhall, and Richard L Scheaffer (2001). *Mathematical Statistics with Applications*. 6th ed. Duxbury Press.
- White, Halbert (1980). “A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity.” In: *Econometrica* 48.4, pp. 817–838.
- (1996). *Estimation, Inference and Specification Analysis*. Econometric Society Monographs. Cambridge: Cambridge University Press.
- (2000). “A Reality Check for Data Snooping.” In: *Econometrica* 68.5, pp. 1097–1126.
- Zakoian, Jean-Michel (1994). “Threshold Heteroskedastic Models.” In: *Journal of Economic Dynamics and Control* 18.5, pp. 931–955.
- Zumbach, Gilles (2007). *The RiskMetrics 2006 methodology*. Tech. rep. RiskMetrics Group.